© Universiti Tun Hussein Onn Malaysia Publisher's Office

IJIE

Journal homepage: http://penerbit.uthm.edu.my/ojs/index.php/ijie
ISSN : 2229-838X e-ISSN : 2600-7916

The International Journal of Integrated Engineering

A Strategic Roadmap for Maximizing Big Data Return

Manal A. Abdel-Fattah [1], Yehia Helmy[2], Thanaa Mohamed Hassan[1,*]

[1]Information Systems Department,
Faculty of Computers and Information, Helwan University, Cairo, 11311, EGYPT

[2] Management Information System Department,
Faculty of Commerce and Business Administration, Helwan University, Cairo, 11311, EGYPT

*Corresponding Author

DOI: https://doi.org/10.30880/ijie.2019.11.07.007
Received 30 May 2019; Accepted 30 July 2019; Available online 10 August 2019

Abstract: Big Data has turned out to be one of the popular expressions in IT the last couple of years. In the current digital period, according to the huge improvement occurring in the web and online world innovations, we are facing a gigantic volume of information. The size of data has expanded significantly with the appearance of today's innovation in numerous segments, for example, assembling, business, and science. Types of information have been changed from structured data-driven databases to data including documents, images, audio, video, and social media contents referred to as unstructured data or Big Data. Consequently, most of the organizations try to invest in the big data technology aiming to get value from their investment. However, the organizations face a challenge to determine their requirements and then the technology that suits their businesses. Different technologies are provided by variety of vendors, each of them can be used, and there is no methodology helping them for choosing and making a right decision. Therefore, the objective of this paper is to construct a roadmap for helping the organizations determine their needs and selecting a suitable technology and applying this conducted proposed roadmap practically on two companies.

Keywords: Big data, Roadmap, Hadoop, MapReduce, Spark, Flink, Cloudera, Hortonworks, MapR

1. Introduction

Big Data need ended up among the hottest research regions today. Looking into the researches of Big Data analysis which could profit businesses, experimental exploration and general society part and also developments done in every sphere. The need is to create framework systems that can explore this Big Data potential to the maximum, without forgetting the challenges associated with its analysis, structure, scale, timeliness and privacy. Companies are confronting the challenge of processing huge chunks of data, and have found that none of the existing unified architectures can proficiently handle this huge volume of data [1].

If we want to have a glance at the data generation history from 1960, we can see this trend in overall: 1960-1990, relational databases; 1990-2000, OLAP technology; 2000- 2010, column-based data storages and cloud computing; and 2010- 2016, Big Data applications. In simple word, the term Big Data means the collection, processing and presenting the massive amounts of data results that come in a variety of formats at high speed [2]. For example, each minute 15h of videos are uploaded to Facebook so that collects more than 50 TB per day. With respect to the amounts of data generating each day, we can predict the growth rate of data in the next years [3].

Companies that need to work with large sets of data have a range of big data, open-source frameworks, tools and solutions from which to choose. Each solution has a different set of advantages, disadvantages and ideal applications which is the aim of this paper is to help these organizations in revealing the selection problems that they fall into when deciding to work with Big Data.

Corresponding author: thanaahassan@hotmail.com

2. Problem Statement

Any company nowadays despite its size deeply thinks it needs to migrate to big data due to increasing technological environment that we live in. However, there is a big confusion towards the ways, the steps and the procedures through which they are going to migrate. Although it isn't hard to argue the value of analyzing big data, it is intimidating to figure out what to do first. For instance, how to start using big data tools? Which vendors to choose from? And if you are already using one of the big data technologies there are yet many questions that go around, if this tool is effective or not? Such undecided companies are lacking the capability of capturing and storing the data that's why they want to adopt a single/package of tools to save time and money but here is the dilemma how to do it? What are the options available? And what is the best choice to select?

3. A Proposed Roadmap For Applying Big data

A strategic roadmap is a plan that defines where a business is, where it wants to go, and how to get it there. It is a visual representation that organizes and presents important information related to future plans. Strategic roadmaps are a common approach to planning. They are an effective communication tool for managers, and link strategic initiatives with business plans. Road mapping acts as a focusing device that marshals efforts toward achieving important goals [4]. Before deciding if your company will be able to work with Big Data or not. A roadmap is conducted in order to help organization in taking their decision as shown in Fig. 1.

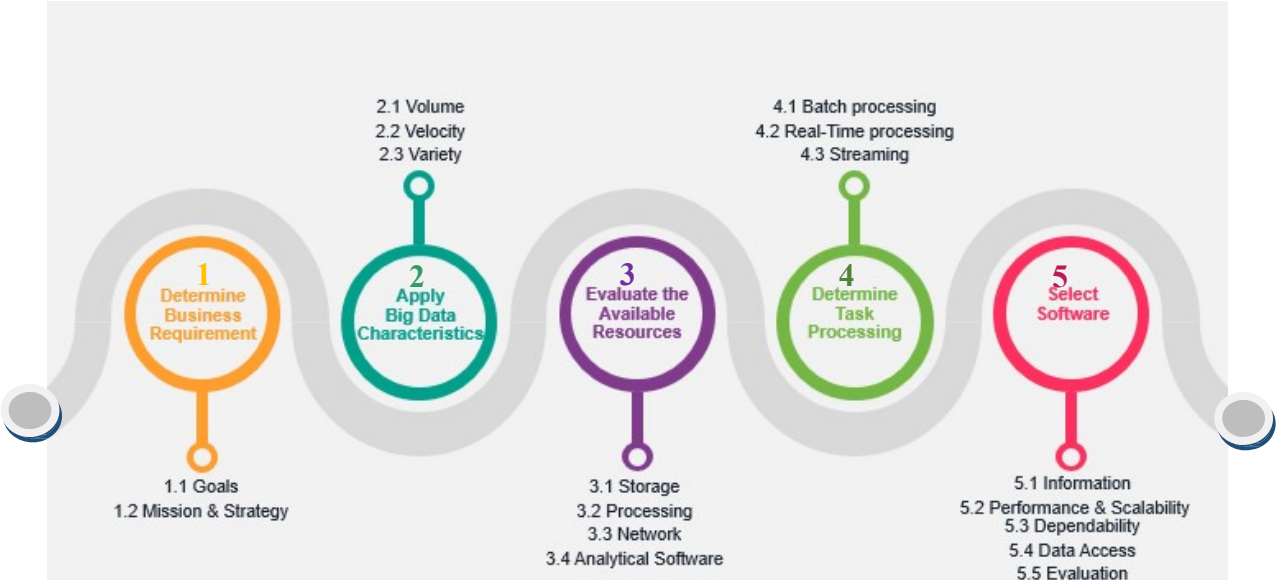

Fig. 1- A proposed Roadmap for applying Big Data (synthesized by the authors)

1. Step 1: Determine Business Requirements

Today, enterprises are exploring big data to discover facts they didn't know before. This is an important task right now because the recent economic recession forced deep changes into most businesses, especially those that depend on mass consumers. In this step we should identify the project business goals, mission and strategy as following [5]:

1.1 Project Business Goals

An eight step approach that any organization should follow in order to define big data project goals [5]:

a) Problem Definition

Determine what the problems you want to solve. In this point, the company need to identify what issues the organization is facing and envision what solutions might be to those problems.

b) Impact

Understand how these problems impact your business and then develop use case(s). Determine: Are you losing millions? Is your staff wasting time by doing more data entry and less analysis? How is this problem impacting your organization?

c) Success criteria
How will you measure the success? What are the top metrics you need to track throughout this process?

d) Value & Impact
What you need to clearly understand is if this problem was solved, what would it mean for your organization? This is typically one of the most crucial steps as it helps determine how and when you should move forward with this project. It also provides context for determining the budget.

If you can't clearly define and articulate steps 1-4, there is no point in moving to step 5.
Also, note that the first 4 steps have little or nothing to do with the technology. This is intentional as you don't want to force technology to solve your business problems. You are starting with the business problem and will map the appropriate technology to solve it.

e) Cloud or On-Premise.
Decide where the solution should live and whether it should be a cloud, on premise, or hybrid solution.

f) Data requirements
Evaluate your data requirement and understand what data is required to solve this problem. Is it data you already have? Is it data you need to go out and get? What is that data and what are the requirements that you need? What is throughput / performance requirements for the data? What are your retention and retrieval requirements?

g) Identify gaps
Determine if this is something that your organization can accomplish with existing or in-house resources and technologies or if you need help from vendors. Do you have enough staff to solve this problem? Are they capable of solving this problem? Will you need additional hardware or software to solve? Identify those gaps and make sure you plan accordingly.

h) Agile or iterative approach.
Start with a pre-production or a pilot implementation. Set goals and milestones and break them up into manageable chunks. Once the pilot is up and running and you see value from it, roll it out into production and enterprise-wide use.

1.2 Project Mission and Strategy

Project Scope: "Scope the project well to deliver near-term business benefit. Using the nucleus project as the foundation for accelerating future big data projects." Since big data projects can get pretty complex, it is helpful to segment the work into broad categories and then drill down into each to create a solid plan [6].

Relevant stakeholders: "The involvement of the relevant stakeholders. It is important that the big data initiative be aligned with key stakeholder requirements. If stakeholders haven't clearly specified their requirements or expectations for your big data initiative, it's not production-ready[6]."

Infrastructure: "In order to successfully support a big data project in the enterprise, you have to make the infrastructure and applications production-ready in your operations." In order to be able to analyze big data (i.e. data is structured and/or not structured at all) you need specific data management technology. One popular big data storage and processing technology that we will talk about is the Hadoop ecosystems of open software tools [6].

Skillsets: "The staff needs to have the right skillsets: e.g. database, integration and analytics skills". This point is not trivial and James Kobielus adds [7]:"Data-driven organizations succeed when all personnel—both technical and business—have a common understanding of the core big data best skills, tools and practices. You need all the skills of data management, integration, modeling, and so forth that you already have running your data marts, warehouses, OLAP cubes".

2. Step 2: Apply Big Data Characteristics

In 2010 Apache Hadoop defined big data as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." As a matter of fact, big data has also been defined as early as 2001. Doug Laney, an analyst of META (presently Gartner) defined challenges and opportunities brought by the increase of data with a 3Vs model, i.e., the increase of Volume, Velocity, and Variety, in a research report [8] and they became the main characteristics of Big Data .

2.1 Volume

Data volume measures the amount of data available to an organization, which does not necessarily have to own all of it as long as it can access it.

2.2 Velocity

Data velocity measures the speed of data creation, streaming, and aggregation.

2.3 Variety

Data variety is a measure of the richness of the data representation – text, images video, audio, etc.

And then those main characteristics were extended to reach 7v's:

a) Value
User can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and can also rank it according to the dimensions they require.

b) Veracity
Big Data veracity ensures that the data used are trusted, authentic and protected from unauthorized access and modification. The data must be secured during the whole their lifecycle from collection from trusted sources to processing on trusted compute facilities and storage on protected and trusted storage facilities [10].

c) Variability
Variability considers the inconsistencies of the data flow. Data loads become challenging to be maintained especially with the increase in usage of the social media which generally causes peak in data loads with certain events occurring [9].

d) Volatility
How old does your data need to be before it is considered irrelevant, historic, or not useful any longer? How long does data need to be kept for?
In a classical data setting, there not might even be data archival policies in place.

We've seen number of companies which realize the most significant benefits from Big Data projects and they are aiming to apply big data in their work in order to provide the highest return and they are in a confusion of how to start. There are two types of companies:

First Scenario: if you are a new company who wants to start working with Big Data, you should have two or more characteristics of these so as to begin using big data that are discussed before in the 7V's model:

i. Encompasses large amounts of information
ii. Consists of a variety of data types and formats
iii. Generated by disparate sources
iv. Retained for long periods
v. Utilized by new and innovative applications

Second Scenario: If you are an existing company that is working with one of the Data Warehousing tools and wants to offload this data to Hadoop you should first understand the meaning of data warehouse and big data keeping in mind that big data isn't a replacement of data warehouse.
In order to well understand the difference between them we should first understand what the two concepts are.

Data warehouse
A Data warehouse is a repository of potentially valuable past and current data that is used to support decision makers [11]. Fig. 2 depicts data warehouse architecture.

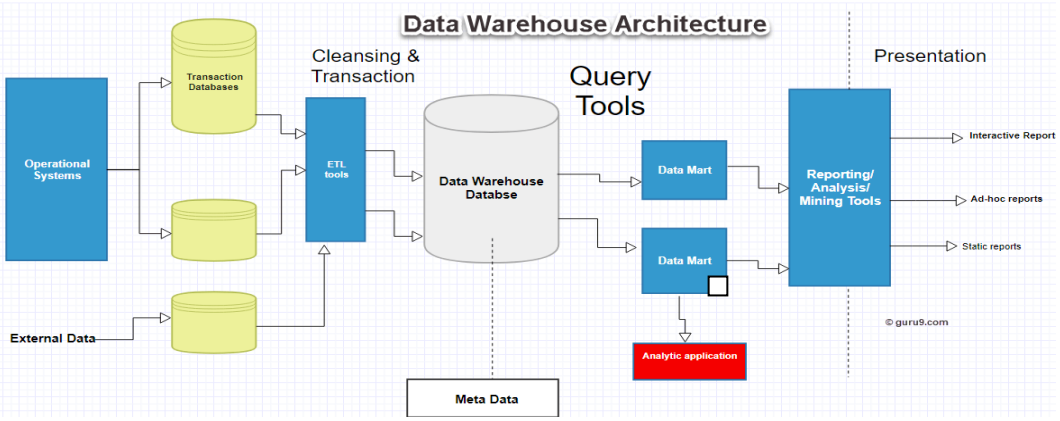

Fig. 2- Data Warehouse Architecture 'adopted from [25]'

Big Data
Big Data is usually linked to unstructured data. This means that there is no relation between pieces of data.
However, Big Data can have any structure. For example, data can be structured, semi-structured, or unstructured. It should never be assumed that Big Data is only unstructured data [11].

Hadoop
When we usually refer to big data we usually refer to Hadoop. Hadoop is a software ecosystem that allows the distribution of large data sets across a cluster of computers [12].

How Data warehouse and Big Data can work together?
Hive is a very interesting Hadoop component as it is a data warehouse tool. Hive provides a SQL like language and a relational model. Hive is a system that sits on top of Hadoop that is designed to process structured data. However, Hive should not be considered a relational database. Hive demonstrates that there are differences between Big Data and data warehousing. Due to the nature of Hive it can be seen certain applications can use both Big Data and data warehousing [11].

To sum up, big data can never replace data warehouse. Big Data is more a technology while the data warehouse is the architect. Technologies like Hive not only demonstrate that Big Data cannot be a replacement for data warehouses but that Big Data and data warehousing can be used together [11].

3. Step 3: Evaluate the Available Resources of Your Company

With the huge quantities of data involved in big data solutions, there must be a robust infrastructure for storage, processing and networking, in addition to analytics software. Nowadays some organizations have the capacity in place to absorb Big Data solutions, while others will need to expand their resources in order to accommodate these new tools, or they could add new capacity to help them in solving the surplus of resources.
Big Data solution won't work properly if one of these 4 infrastructure solutions is missing:

3.1 Storage

Often, organizations already possess enough storage in-house to support a Big Data initiative. The initial test server for Hadoop cluster is:

- 1 Name Node (64GB ram + 24 core) + 2 Hard Disk Drive (1 for OS, 1 for HDFS storage).
- 3 Data Node (each 32GB ram + 16 core) + 2 Hard Disk Drive (1 for OS, 1 for HDFS storage).Data Node is also used for: zookeeper, Kafka, spark, YARN/MapReduce, Impala and Pig/Hive gateway.

As the best practice to run Hadoop environment, all server should be a bare metal.

3.2 Processing

Servers intended for Big Data analytics must have enough processing power to support the Big Data application. Some analytics vendors, such as Splunk (Splunk Analytics for Hadoop is an integrated analytics add-on that enables anyone in your organization to interactively explore, analyze and visualize raw, unstructured big data), offer cloud processing options, which can be especially attractive to agencies that experience seasonal peaks [13].But if you are a daily demanding company you can rely on on-premises processing resources to handle the steadier, day-to-day demands.

3.3 Networks

The massive quantities of information that must be transferred back and forth in a Big Data initiative require robust networking hardware.

Many organizations are already operating with networking hardware that facilitates 10-gigabit connections, and may have to make only minor modifications — such as the installation of new ports — to accommodate a Big Data initiative. Securing network transports is an essential step in any upgrade, especially for traffic that crosses network boundaries [13].

3.4 Analytics Software

Agencies must select Big Data analytics products based not only on what functions the software can complete, but also on factors such as data security and ease of use.

One popular function of Big Data analytics software is predictive analytics — the analysis of current data to make predictions about the future. Predictive analytics are already used across a number of fields, including actuarial science, marketing and financial services. We can also find Governmental applications include fraud detection, capacity planning and child protection [13].

4. Step 4: Determine Task Processing

4.1 Batch Processing

Batch processing main focus is on banning any computation of data unless all the input data is completely available, so the Hadoop framework processes the input data and the output results will be available when all of the computation is done.

4.2 Real-Time Data

In big data processing, real-time queries always respond in terms of seconds and minutes instead of batch jobs which takes hours and days instead.

4.3 Streaming Data

Today's applications need more stream-like demands in which the input data is not available completely at the beginning and arrives constantly.

MapReduce, Spark and Flink are one of the most popular Big Data task processing frameworks. Each one of them has weaknesses and strengths and choosing between them is based on demands and applications. Table1 is conducted between those popular frameworks as following [14] [15] [16] [17]:

Table 1- A Comparison between MapReduce, Spark and Flink

	Hadoop	**Spark**	**Flink**
Data Processing	Batch processing system	Real Time processing system but it also supports streaming processing system but not adequate in case of complex streaming	Provides single runtime for the streaming and batch processing
Processing Speed	MapReduce processes slower than Spark and Flink	Spark is 100 times faster than MapReduce because of its memory processing system	Faster than spark because of its streaming engine
Optimization	Apache MapReduce jobs has to be manually optimized	Apache Spark jobs has to be manually optimized	Apache Flink jobs has to be automatically optimized
Scheduler	MapReduce needs an external job scheduler like **Oozie** in order to perform complex jobs	Due to in-memory computation, spark acts its own flow scheduler	Flink can use YARN Scheduler but Flink also has its own Scheduler
Caching	MapReduce cannot cache the data in memory for future requirements	It can cache data in memory for further iterations which enhance its performance	It can cache data in memory for further iterations which enhance its performance
Cost	MapReduce can typically run on less expensive hardware than some alternatives since it does not attempt to store everything in memory	Spark requires a lot of RAM to run in-memory, increasing it in the cluster, gradually increases its cost	Apache Flink also requires a lot of RAM to run in-memory, so it will increase its cost gradually
Easy to use	MapReduce developers need to hand code each operation which makes it very difficult to work	It is easy to program as it has tons of high-level operators	It also has high-level operators
Compatibility	Apache Hadoop MapReduce and Apache Spark are compatible with each other	Apache Spark and Hadoop are compatible to each other	Apache Flink is a scalable data analytics framework that is fully compatible to Hadoop

5. Step 5: Software Selection

We get asked a lot of questions about how to select Apache Hadoop. Remember that Hadoop is built to handle component failure well and to scale out on low cost gear. So when looking about choosing the best Hadoop company, you will find a lot of them as following:

- MapR
  MapR Platform delivers enterprise grade security, reliability, and real-time performance while dramatically lowering both hardware and operational costs of your most important applications and data [18].

- IBM
  IBM solves this challenge with a zone architecture optimized for big data. The next generation architecture for big data and analytics delivers new business insights while significantly reducing storage and maintenance costs.
  IBM and Hortonworks have partnered to give enterprises easy access to the capabilities, scalability and economy of Apache Hadoop, plus additional governance and security features as well as tools for data federation, advanced query and management of their data [19].

- Amazon Web Service
  Amazon Web Services offers a broad set of global cloud-based products including compute, storage, databases, analytics, networking, mobile, developer tools, management tools, IoT, security and enterprise applications. These services help organizations move faster, lower IT costs, and scale.
  Amazon Web Service had made a joint solution with Cloudera's expertise in large-scale data management and analytics with AWS' expertise in cloud computing [20].

- Hortonworks Data Platform
  HDP is the industry's only true secure, enterprise-ready open source Apache Hadoop distribution based on a centralized architecture (YARN). HDP addresses the complete needs of data-at-rest, powers real-time customer applications and delivers robust analytics that accelerate decision making and innovation [22].

- Cloudera
  Cloudera Enterprise is the fastest, easiest, and most secure modern data platform. From analytics to data science, anyone can now get results from any data and across any environment—all within a single, scalable platform [21].

- Teradata

  Teradata helps companies get more value from data than any other company. The big data analytic solutions and team of experts can help your company gain a sustainable competitive advantage with data
  Hortonworks and Teradata believe that no single analytic system can meet all customers' needs, and leading data driven organizations will deploy an analytical ecosystem.

  That is why Hortonworks and Teradata have a longstanding engineering partnership focused on delivering the analytical ecosystem to the enterprise through best in class analytic engines and co-engineering of software to orchestrate the processing and strip out the complexity typically associated with a hybrid environment [23].

Comparing Major Hadoop Distributions

Every enterprise is searching for ways to profit from Big Data. Which means that Hadoop is likely to have a major role to play in your organization. Given this probability, you should take your consideration well of the Hadoop implementation, and pay all the attention to its performance/scalability, dependability, and ease of data access. In particular, make sure that you have selected the best Hadoop that will fit that way you operate, and not the other way around.
So as to take a decision in choosing the best Hadoop distributor. A quick comparison chart is conducted between the biggest three major Hadoop distributors companies (Cloudera, Hortonworks, MAPR) as shown in table2 [24].

Table 2- A Comparison between MapR, Hortonworks and Cloudera

	Cloudera	Hortonworks	MAPR
1. General Information			
Founded Year	2009	2011	2009
License	Multiple versions: open source and licensed	Open source	Licensed
Execution Environment	Local or Cloud	Local or Cloud	Local or Cloud
Management/administration Tools	Cloudera Manager	Ambari	MapR Control System
Free Trial	Available	Available	Available
Support Services	Available	Available	Available
Base	HDFS , YARN	HDFS , YARN	HDFS , YARN
Solution Components	Hive, Pig, HBase, Hue, Avro, Whirr,	Hive, Pig, HBase, Hue, Tez, Yam,	Tez, Spark, Cascading, Pig, MapReduce,

	Flume, Yarn, Mahout, Cloudera Impala, Cloudera Manager, Sqoop, Oozie, Zookeeper, Apache Sentry	Zookeeper, Mahout, Flume, Sqoop, Oozie, Whirr, Storm, Apache Ganglia, Apache Falcon, WebHDFS, NFS, Spark, Accumulo, Knox	GraphX, MLLib, Mahout, Drill, Shark, Impala, Hive, Accumulo, Solr, HBase, Storm, Hue, Flume, Sqoop, Knox, Sentry, Falcon, Oozie, Whirr, Zookeeper.
2. Performance and Scalability			
Data Ingest	Batch	Batch	Batch and streaming rights
Metadata Architecture	Centralized	Centralized	Distributed
HBase Performance	Latency spikes	Latency spikes	Consistent low latency
NoSQL Application	Batch application	Batch application	Batch and online application
3. Dependability			
High Availability	Single failure recovery	Single failure recovery	Self-healing across multiple failure
MapReduce HA	Restart jobs	Restart jobs	Continuous without restart
Upgrading Replication Snapshots	Rolling upgrades Data Consistent only for closed files	Planned downtime Data Consistent only for closed files	Rolling upgrades Data + Metadata Point in time consistency for files and tables
Management Tools	Cloudera Manager	Ambari	MAPR control system
Volume Support	No	No	Yes
Alarms, Alerts	Yes	Yes	Yes
Data and Job placement job	No	No	Yes
4. Data Access			
File system access	HDFS, read-only NFS	HDFS, read-only NFS	HDFS, read/write NFS
File I/O	Append only	Append only	Read/Write
Security: ACLs	Yes	Yes	Yes
Wire-level Authentication	Kerberos	Kerberos	Kerberos, Native
5. Evaluation			
Strength Point	Cloudera has a user friendly interface with many features and useful tools like Cloudera Impala	It is the only Hadoop Distribution that supports Windows platform	It is one of the fastest Hadoop distribution with multi node direct access
Weakness Point	Cloudera is comparatively slower than MapR Hadoop Distribution	The Ambari Management interface on HDP is just a basic one and does not have many rich features.	MapR does not have a good interface console as Cloudera
Conclusion	If the company seeks somewhere in between open source platform and the product features and management innovation then Cloudera is the best choice	If the company seeks open source platform then Hortonworks is the best choice	If the company seeks the feature of the product itself then MapR is the best choice

4. Case Study 1: Commercial International Bank (CIB)

Background

Big data and advanced analytics have been playing a vital and important role in transforming the way business is done as well as influencing the behavior of customers in all walks of life, and the banking industry is no exception. To implement these rapid developments, bold and swift action is required in order to lead change in the industry.
Data has evolved from being just a tool to become a commodity in itself. Speaking of data, it represents a fundamental shift in how business is done by opening a door to a new strategy approach [26].
That's what CIB as a bank decided to do, they decided to choose big data analytics in transforming their business. Big data analytics will enable a colossal opportunity for CIB by giving them a clear edge to lead the market. Banks that can have the opportunity to transform enormous amounts of data into valuable insights and then actions will be able to create a unique and differentiated customer experience [26].

Reasons for CIB Big Data Transformation

CIB is Egypt's leading private sector bank was founded in 1975, offering a comprehensive and wide range of financial products and services to its clients, who include enterprises and institutions of all sizes, high-net-worth (HNW) individuals and retail customers [26]. CIB operates in every banking sector segment including corporate, commercial, retail wealth management and SME, all of this delivered through client-centric teams. Therefore, they decide to invest aggressively in IT and Human Capital in order to develop exceptional infrastructure that can support them through their huge journey [26].

Apply Big Data Proposed Strategic Road-map

In 2016, CIB decided to conduct Big Data analytics and had decided to have that done with the help of Teradata Company.

Step1: Determine Business Requirements

i. CIB Goals

CIB understand the importance of moving from a descriptive analytics model to predictive analytics and fully comprehend the challenges and difficulties this entails. Therefore, they decide to invest aggressively in IT and Human Capital to develop exceptional infrastructure that can support them through this journey.

CIB invested in data storage and computation platforms in order to [26]:

- Increase structured data capacity
- Improve reporting performance
- Invest in self-service business intelligence
- Invest in real-time information delivery systems to manage petabytes of data for advanced analytics and new regulatory requirements.

ii. CIB Mission and Scope

CIB Scope

CIB scope is to uphold its distinct reputation as one of the leading and trusted financial institution in Egypt, respected for its people, strong core values, performance, and commitment to inclusive, responsible, and sustainable growth and as a bank who is always seeking to be the first in using any updated technologies in order to better enhance their institution .

CIB Mission

Speaking of CIB Mission, they are always creating an outstanding stakeholder value by providing best-in-class financial solutions to the individuals and enterprises that drive Egypt's economy. This is done through innovative product offerings, superior customer service, staff development strategies, and commitment to sustainability. As a bank, they are confident that their ambitions will be fulfilled to help shape the future of banking in Egypt for years to come.

Step 2: Apply Big Data Characteristics

CIB should have big data main characteristics which are Volume, Velocity and Variety.

- Volume: CIB amount of data aren't that big and all of them are structured and can be stored easily without any need for Hadoop.
- Velocity: The Bandwidth and data ingestion for measuring data creation, streaming, and aggregation can be easily done with data warehouse tools.

- Variety: No variety of data representation from text, images, audios and videos that is in need for applying big data technology.

Table 3- Applying Big Data Characteristics on CIB

Big data Characteristics	CIB
Encompasses large amounts of information	✓
Consists of a variety of data types and formats	✓
Generated by disparate sources	✓
Retained for long periods	YES
Utilized by new and innovative applications	YES

CIB Case Conclusion:
Although moving forward in technology is very important and vital for all big corporation. Every organization had his need that suits it more in order to create more value with their resources. CIB in this state, the storage technology needed for it was data warehouse storage more than the HDFS (Hadoop Distributed File System) with applying some analytical and predictive tools.

5. Case Study 2: Cisco

Background

Cisco is one of the worldwide leader in networking that help in the process of transformation of how people connect, communicate and collaborate. Cisco IT manages of about 38 global data centers comprising 334,000 square feet. Approximately 85 percent of applications in newer data centers are virtualized and IT is working in make a huge difference by increasing this percentage toward a goal of 95 percent virtualization. At Cisco, very large datasets about customers, products, and network activity is representing hidden business intelligence. The same is true of unstructured terabytes of data such as web logs, video, email, documents, and images [27].

Cisco decided to work on an Enterprise Hadoop architecture, built on Cisco Unified Computing System (UCS) Common Platform Architecture (CPA) for Big Data but the question here what made them capable to make this move.

Reason for Cisco Big Data Transformation

- Exponential increases in data volumes which means that big data environments will eventually get much bigger and more distributed, potentially consisting of hundreds or thousands of servers and switches.
- IT departments need an infrastructure which will be designed for cost-effective management, massive scalability, comprehensive security that extends from the data center through the network to every connected device, and lower TCO.
- Big data had become critical for day-to-day decision making and more pervasive in the industry where big data solutions help in ensuring faster and predictable delivery of insights to key decision makers.
- Traditional infrastructure simply is no longer can handle these long-term requirements of big data environments.

Apply Big Data Proposed Strategic Road-map

Step 1: Determine Cisco Business Requirements

i. Cisco Goals

Unlock the large data sets business value that includes structured and unstructured information.
Provide service-level agreements (SLAs) for internal customers who are using big data analytics services.
Provides a support for multiple internal users on same platform [28].

ii. Cisco Mission & Strategy

- Implement an enterprise Hadoop platform on Cisco UCS Common Platform Architecture (CPA) for Big Data that is described as a complete infrastructure solution including compute, storage, connectivity and unified management.
- Enables an automated job scheduling and process orchestration using Cisco Tidal Enterprise Scheduler as alternative to Oozie [28].

iii. Cisco Skillset needed

Cisco IT server administrators responsibility will be the managing of all elements of the Cisco UCS including servers, storage access, networking, and virtualization from a single Cisco UCS Manager interface. Cisco UCS

Manager will help us in managing larger clusters as our platform grows without any need for increasing staffing this will help in saving time and effort also eliminated configuration errors that could cause downtime.

Step 2: Apply Big Data Characteristics

Cisco as a company had all the characteristics needed to apply Big Data as following:

- Volume: Cisco encompasses large amounts of information about customers, products and networks activity representing hidden business intelligence.
- Variety: Cisco consists of a variety of data types and formats such as web logs, videos, email, documents, and images.
- Velocity: Cisco had a high speed in data creation, streaming and aggregation generated by disparate sources retained for long periods.

Step 3: Evaluate the available resources of Cisco

Cisco UCS CPA for Big Data provides the capabilities we need in order to use big data analytics for business advantage, including high-performance, scalability, and ease of management," says Jag Kahlon, Cisco IT architect [27].

i. Storage

Linearly Scalable Hardware with Very Large Onboard Storage Capacity The compute building block of the Cisco IT Hadoop Platform is the Cisco UCS C240 M3 Rack Servers, 2-RU (Rack Unite) server that is powered by two Intel Xeon E5-2600 series processors, 256 GB of RAM, and 24 TB of local storage. Out of the 24 TB, Hadoop Distributed File System (HDFS) will be able to use 22 TB, and the remaining 2 TB will be available for the operating system [27].

ii. Processing

Cisco UCS C-Series Servers helps in providing a high performance access to local storage which is the biggest factor in Hadoop. The current architecture comprises and consists of four racks, each containing 16 server nodes supporting 384 TB of raw storage per rack. "This configuration can be able to scale 160 servers in a single management domain supporting 3.8 petabytes of raw storage capacity," says Kahlon [27].

iii. Network

Low-Latency, Lossless Network Connectivity Cisco UCS 6200 Series Fabric Interconnects provides high speed, low latency connectivity for servers and centralized management for all connected devices by the UCS Manager. Deployed in redundant pairs offers the full redundancy, performance (active-active), and exceptional scalability for large number of nodes typical in clusters of big data. Each rack connects to the fabric interconnects through a redundant pair of Cisco Nexus® 2232PP Fabric Extenders, which behave and acts like a remote line cards [27].

Step 4: Determine Task Processing

i. Batch processing

Cisco hadoop is a strategic data platform embraced by batch streaming enterprises which is MapReduce. MapReduce is well-known as a distributed computing framework. Data is processed on the same Cisco UCS server where it resides as well as avoid latency during the access of data over the network [28].

ii. Real time data processing

Real-Time Systems also need to process the data in addition to guarantee the response within specific time constraints, and return all the results that will affect the environment they are running in. The powerful, easy-to-use for these use cases is the open source platform Apache Spark. Apache Spark enables in-memory capabilities, it offers both real-time and batch processing capabilities over a wide range of scenarios. Speaking of Cisco, it is in need of Real time data processing which involves a continual input, process and output of data. Data must be processed in a small time interval (or near real time). Real time data processing and analytics provides any organization with the ability to take immediate action for those times when acting within seconds or minutes is significant. The main goal is to obtain the insight required and act prudently all of that within the right time [27].

Step 5: Cisco Software Selection

i. Information

Cisco IT uses MapR Distribution for Apache Hadoop, which helps in speeding up MapReduce jobs with an optimized shuffle algorithm, direct access to the disk as well as built-in compression. The Cisco UCS CPA with MapR solution together delivers a powerful and flexible infrastructure that helps in increasing both business and IT agility, reducing the total cost of ownership (TCO), and delivering an exceptional return on investment (ROI) at scale, while fundamentally transforming the way organizations perform business with Hadoop technology [29].

ii. Performance and Scalability

MapR data ingest performance is both batch and streaming, the HBase performance is consistent low latency and the NoSQL application is batch and online application.

iii. Dependability

MapR provides self-healing across multiple failure, their MapReduce is continuous without restart. MapR replicates both data and metadata snapping shots for files and tables at point in time consistency. MapR provides volume support, alarms, alerts, data and job placement job.

iv. Data Access

MapR file system access for HDFS is read and write and also the file I/O read and write. The MapR wire level authentication is Kerberos, Native.

v. Evaluation

The Cisco UCS CPA in its combination with MapR solution is based also on the Cisco® Common Platform Architecture for Big Data. Speaking of its fabric-based infrastructure, Cisco UCS CPA for Big Data delivers an exceptional performance, capacity, management simplicity, and also scale to help customers derive value from the biggest data challenging deployments [29].

Table 4 - Software Selection

	Cloudera	Hortonworks	MAPR
Performance and Scalability			
Batch and Streaming Data Ingest	Batch only	Batch only	✓
Distributed Metadata Architecture			✓
HBase Performance	Latency spikes	Latency spikes	Consistent low latency
NoSQL Application	Batch application	Batch application	Batch and online application
Dependability			
Self-healing multiple failure			✓
Mapreduce HA	Restart jobs	Restart jobs	Continuous without restart
Upgrading	Rolling upgrades	Planned downtime	Rolling upgrades
Data and Metadata Replication			
Manageability			
Volume Support			✓
Alarms, Alerts	✓	✓	✓
Data and Job placement job			✓
Data Access			
HDFS, read-write File system access			✓
File I/O	Append only	Append only	Read/Write
Security: ACLs	✓	✓	✓
Wire-level Authentication	Kerberos	Kerberos	Kerberos, Native

Cisco Conclusion

Speaking of the combination between MapR and Cisco UCS CPA, it brings the power of MapR to a dependable deployment model which you will be able to implement rapidly and customize for either high performance or high capacity using Cisco Unified Fabric and powerful and efficient Cisco UCS rack servers. Whether you are deploying a large data center or you are just buying single racks through the Cisco Smart Play program, you can size the Cisco UCS CPA with MapR solution in order to meet the challenges of Hadoop.

6. Conclusion

This paper discusses Big Data in order to provide a brief overview of this new trend, unveiling a vast world of technologies which are impossible to discuss in this paper due to space limitations. A proposed roadmap has been conducted in order to help any enterprise to take the appropriate decision when starting working with big data tool. This roadmap consists of 5 steps first you have to determine your organization business requirements of goals and mission, then make sure that your company acquire big data characteristics (Volume, Velocity, Variety) after the second step (if you can't clearly define and articulate steps 1 & 2 , there is no point in moving to step 3),evaluate also your enterprise resources in order to apply big data software from processor, storage and network, after that any organization should determine what will be the most appropriate task processing type that suits their organization and last a conducted comparison was done to help this enterprise choose and select the best Hadoop software that suits their organization. We conducted this roadmap one two company's one of them in the banking sector which is CIB. CIB had found that they are not in need to apply Hadoop and they had stopped on step 2 in the proposed roadmap. The other company was Cisco that had moved on all the 5 steps, selecting at the end the MapR Hadoop distributor. As future work, we pretend to intensify our research in Big Data in order to be able to host and respond to high demand Big Data solutions.

References

[1] Hussain, T., & Farah, A. Big Data-Tools and Technologies.

[2] Mohd Sharef, N., M. Shafazand, Y., Ahmad Nazri, M. Z., & Husin, N. A. (2018). Self-adaptive Based Model for Ambiguity Resolution of The Linked Data Query for Big Data Analytics. International Journal of Integrated Engineering, 10(6).

[3] Gheisari, M., Wang, G., & Bhuiyan, M. Z. A. (2017). A survey on deep learning in big data. In Computational Science and Engineering (CSE) and Embedded and Ubiquitous Computing (EUC), 2017 IEEE International Conference, 173-180.

[4] Chen, J., Chen, Y., Du, X., Li, C., Lu, J., Zhao, S., & Zhou, X. (2013). Big data challenge: a data management perspective. Frontiers of Computer Science, 7(2), 157–164.

[5] Azcentral. (2013). What Is a Strategic Roadmap? Retrieved on August 29, 2018 from: https://yourbusiness.azcentral.com/strategic-roadmap-14662.html.

[6] IKANOW Homepage. Retrieved on July 29, 2018 from http://www.ikanow.com/8-proven-steps-to-starting-a-big-data-analytics-project/

[7] Zicari, R. V., Rosselli, M., Ivanov, T., & Kor, N. (2016). Setting Up a Big Data Project : Challenges , Opportunities , Technologies and Optimization. Big Data Optimization: Recent Developments and Challenges, 18, 17-47.

[8] Kobielus, J. (2015). How to do a big data project: A template for success. Infochimps White Paper, 1–17.

[9] Shi, Y. et al., (2001). Early endothelial progenitor cells as a source of myeloid cells to improve the pre-vascularisation of bone constructs. Europian Cells & Materials, 27, 64–80.

[10] Katal, A., Wazid, M., & Goudar, R. H. (2013). Big data: issues, challenges, tools and good practices. In Contemporary Computing (IC3), IEEE Sixth International Conference, 404-409.

[11] Demchenko, Y., Grosso, P., De Laat, C., & Membrey, P. (2013). Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), IEEE International Conference, 48-55.

[12] Akhtar, N., Saleh, J. M., & Grelck, C. (2018). Parallel Processing of Image Segmentation Data Using Hadoop. International Journal of Integrated Engineering, 10(1).

[13] Apache Hadoop. Retrieved on August 6, 2018 from: http://hadoop.apache.org

[14] W. Paper and E. (2015). Summary by Federal Data Analaytics, Making Sense Of Big Data.

[15] Marcu, O. C., Costan, A., Antoniu, G., & Pérez-Hernández, M. S. (2016). Spark versus flink: Understanding performance in big data analytics frameworks. In Cluster Computing (CLUSTER), 2016 IEEE International Conference, 433-442.

[16] Landset, S., Khoshgoftaar, T. M., Richter, A. N., & Hasanin, T. (2015). A survey of open source tools for machine learning with big data in the Hadoop ecosystem. Journal of Big Data, 2(1), 24.

[17] García-Gil, D., Ramírez-Gallego, S., García, S., & Herrera, F. (2017). A comparison on scalability for batch big data processing on Apache Spark and Apache Flink. Big Data Analytics, 2(1), 1.

[18] Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache flink: Stream and batch processing in a single engine. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 36(4).

[19] Reifer, A. (2016). Inside the MapR Hadoop distribution for managing big data. Retrieved on August 11, 2018 from: https://searchdatamanagement.techtarget.com/feature/Inside-the-MapR-Hadoop-distribution-for-managing-big-data

[20] Whitepaper:Hortonworks. (2018). The Power of One: IBM + Hortonworks. Retrieved from https://public.dhe.ibm.com/common/ssi/ecm/27/en/27016927usen/hybrid-cloud-hybrid-cloud-white-paper-external-27016927usen-20180607.pdf

[21] Amazon Web Services. (2018). Apache Hadoop on Amazon Web Services (AWS). Retrieved on August 15, 2018 from: https://aws.amazon.com/emr/details/hadoop/

[22] Cloudera. (2018). (White Paper). Cloudera Enterprise Reference Architecture for AWS Deployments. Retrieved from: http://www.cloudera.com/documentation/other/reference-architecture/PDF/cloudera_ref_arch_aws.pdf

[23] PAT Research. (2017). Hortonworks Data Platform - Compare Reviews, Features, Pricing in 2018 -PAT RESEARCH: B2B Reviews, Buying Guides & Best Practices. Retrieved from https://www.predictiveanalyticstoday.com/hortonworks-data-platform/

[24] Hortonworks. (n.d.). Teradata - Hortonworks. Retrieved from https://hortonworks.com/partner/teradata (accessed on 29 August 2018)

[25] Schneider, R. D. (2015). Hadoop Buyer's Guide. Ubuntu.

[26] Data Warehouse Concepts, Architecture and Components. (n.d.). Retrieved from https://www.guru99.com/data-warehouse-architecture.html (accessed on 31 August 2018)

[26] CIB Egypt Website. [cited 2018 26 September]; Available from: https://www.cibeg.com.

[27] Cisco Corporate Information. [cited 2018 28 September]; Available from: https://www.cisco.com/c/dam/en_us/about/ac49/ac20/downloads/annualreport/ar2008/pdf/cisco_ar2008_corporate_information.pdf.

[28] How Cisco IT Built Big Data Platform Using MapR Distribution to Transform Data Management. [cited 2018 28 September]; Available from: https://mapr.com/resources/how-cisco-it-built-big-data-platform-using-mapr-distribution-transform-data-management/.

[29] Cisco and MapR Joint Brochure. 2016 [cited 2018 1 October]; Available from: https://www.cisco.com/c/dam/assets/docs/ucs-and-mapr.pdf