Big Data: Trends, Tools & Technologies, Open Issues and Opportunities

Author Name: Chetan Khatri

Feb, 2015

chetan@kutchuni.edu.in

http://cs.kutchuni.edu.in/

Department of Computer Science
KSKV Kachchh University, Bhuj.
Gujarat - INDIA

Contents

Chapter 1
Big Data

1.1 What is Big Data?

Big data is the data which is expensive to extract, transform, load for decision making in an Enterprise, challenges include capture, duration, storage, search, sharing, transfer, analysis and visualization.

"Big Data is any data that is expensive to manage and hard to extract value from"
-Michael Franklin (University of Barkley)

"Big data, have entered into world beyond just static data that collected" - TDWI

"Big data is overwhelming of data, this data has challenges volume of data, unstructured way of data, confidentiality" - Adobe at TDWI

"Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, and everyone thinks everyone else is doing it, so everyone claims they are doing it."
-Dan Ariely

"I'm a data janitor. That's the sexiest job of the 21st century. It's very flattering, but it's also a little baffling"
- Josh Wills, a senior director of data science at Cloudera

"Given enough data, everything is statistically significant" - Douglas Merrill

1.2 5'V and C Explanation

Volume: The amount / size of data available for processing. The characteristic notify whether data is big data or not.
Variety: the different formats of data in enterprise like sensor data, flat file, xml file, documents , binary data, Relational database etc.
Velocity: the speed in which data has been generated and processed.
Veracity: the quality and accuracy of data, whether that contains missing or noisy values.
Variability: the inconstancy shown by data at times, thus hampering the process of being able to handle and manage the data effectively.
Complexity: to process the entire data is complex task as concern with Big data, it should be connected, it should be correlated for accurate decision making, complexity might face while pre-processing data.

1.3 What is Data Science?

Data Science is the field where data is acquires, processes, manipulates, data science is combination of Engineering, Mathematics, data warehouse, data mining, database management system, Machine Learning, Artificial Intelligence, Algorithm, Programming, and statistics.

The phenomenon in technology development significantly exposes the staggering growth of data, as much as growth of data goes much and much more, data science skill requires more to handle that growth and scale of new types of data from sensor, social media, website logs, click steaming etc.

In other words, data science can be broken down into four essential parts.
Mining Data: Collecting and formatting various types of data based on pattern mechanism.
Statistics: Gathered information must be analyzed.
Interpret: Representation or visualization in the form of Presentation, charts, graphs, reports.
Leverage: Studying Implication of the data, application of data, tools & technologies of data, Interaction and prediction of data.

1.4 Differentiate Big data and Data Science

Data science is field where areas like Engineering, Mathematics, data warehouse, data mining, database management system, Machine Learning, Artificial Intelligence, Algorithm, Programming, and statistics included. Where big data have modern applications and technologies to manage and process those data, Big data includes data sets whose size and type make them impractical to process and analyze with traditional database technologies.

Data science is very impressive field in 21st century, the person who knows above skills is known as "Data Scientist"

Data Scientist defined as " A good Scientist understand importance of:

- Their eyes search for Information on the web.
- Algorithmic Strategizing.
- Verctorized operations.
- Have knowledge of latest tools and technologies to handle data.
- Efficient in data mining, statistics, mathematics, artificial intelligence.

"I keep saying that the sexy job in the next 10 years will be statisticians. And I'm not kidding." - HAL VARIAN, chief economist at Google.

1.5 Collection of data / Filtering /Classification of Data

Internet of Things (IoT) generates different types of data very rapidly such as from sensors of car, satellite, air craft, health care, weather, maps etc. All data are in different form that is real problem big data has to store, process, transform, and analyze different types of data in row format. As per characteristics of data we are categorized in following 3 types.

There are 3 basic types of Data:

- Structured Data
- Unstructured data
- Semi structured data

Structured Data:

Structured data is usually data from RDBMS that are arranged in well defined manner, so each and every defined field has own unique value and description, and fields are relative defined with another field. So it's already known that which values goes where.

Unstructured data:

Unstructured data is usually the data don't have relational structure, today rapid growth unstructured data is 85 % of all data in world, unstructured data doesn't make any sense without pre-processing it, basically documents, text, presentation, web logs, sensor signal data is best examples of unstructured data.

Unstructured data either generated by human or machine generated.

Human generated unstructured data such as, document, presentation, web page, text file etc.

Machine generated unstructured data belongs to following categories:

- Satellite image: this includes weather data or government capture digital surveillance imaginary. Ex. Google earth.
- Scientific data: This includes seismic imagery, atmospheric data, and high energy physics.
- Photographs and video: This includes security, surveillance, and traffic video.
- Radar or sonar data: This includes vehicular, meteorological, and oceanographic seismic profiles.
- Aircraft data: Flying aircraft generates large amount of unstructured data, that is consumed by Air traffic controller.

Semi- Structured data:
Semi-Structured data is a form of structured data that doesn' t have relational database or table format, but it has tags to separate different entities and sub-entities, therefore it's known as self-describing structure.
In Semi-structured data, the entities belonging to the same class may have different attributes even though they grouped together, and the attributes' order is not important at all.
Semi-Structured data is increasingly occurring since the advent of the Internet where full-text documents and databases are not the only forms of data any more and different application requires medium to exchange information, in object - oriented database you finds semi- structured data.
Types of Semi-Structured data:
1. XML
2. JSON

1. **XML:** xml is user-defined mark-up language where user can generate own hierarchy for their data, XML is being popularize by SOAP based web services. Ex. WSDL
   XML has human readable data structure, however, can only be taken so far. Xml data structure is human readable but sometimes it requires domain knowledge otherwise it looks like American native child reads Swahili!
2. **JSON:** JSON stands for JavaScript Object notation, this is open standard to do data transmission with attributes in key-values pairs, and Restful web service consumes and produces JSON types of data, JSON widely used in client - server data communication. Nowadays, most of server produces JSON data for their client-applications. Ex. Twitter API, Facebook graph API etc.

1.6 Big data life cycle [1]

Organization is eager to harness the power of big data, but as new big data opportunities emerge, ensuring that information is trusted and protected becomes exponentially more difficult, If this technologies is not implemented in proper stages then organization may failed to reduce Volume, Variety, Velocity issues in big data.
The tremendous volume, variety and velocity of big data mean that the old manual methods of discovering, governing and correcting data are no longer feasible. Organizations need to automate information integration and governance from the start.
Big data lifecycle can help organization to protect, secure, and improve the accuracy of data. Information integration and governance solutions must become a natural part of big data projects. They must support automated discovery and profiling and they must facilitate an understanding of diverse data sets to provide the complete context required to make informed decisions.
They must be agile enough to accommodate a wide variety of data and seamlessly integrate with diverse technologies, from data marts to Apache Hadoop systems. Plus, they must discover, protect and monitor sensitive information across its lifecycle as part of big data applications. Understanding the context of data and being able to extract the precise information necessary to meet a business Objective is key to utilizing big data to the fullest. Managing the data lifecycle so that data is accurate, is appropriately used and is correctly stored to meet the required service levels and retention needs has wide-ranging benefits. These benefits include risk reduction, performance improvements and preventing an overload of useless information.

Without effective data lifecycle management, the increasing Volume, Variety, Velocity, Veracity, Complexity can reduce the performance, and increase the margin and risk factor.
Performance
Based on growth of Big data issues such as Volume, Variety, Velocity, Veracity, Complexity. More and more queries executes on data takes much response time will increase the entire

performance of the project. If unchecked, continued data growth will stretch resources beyond capacity and negatively impact on resources beyond the capacity and negatively impact response time for critical queries and reporting processes, these problems can affect production environment as well as hamper to archiving, migration and disaster recovering the efforts. Increasing of data can also affect testing data because the Exabyte's of data processing requires longer and repeating data life cycle execution.

Margin

Exponential data growth can also increase infrastructure and operational cost, rising data volume requires high capacity server, hardware and also power that can increase cost to business.

Rising of data volume also requires monitoring the health and performance of the infrastructure to monitoring resources such as servers and database engine services, it also requires several of software applications and it cost to buying license for performance monitoring tools.

Risk

Following the "let's keep it in case someone needs it later" mandate, many organizations already keep too much historical data. According to the CGOC 2012 Summit Survey, 69 percent of data has no value. Opening the doors to excessive storage and retention only exacerbates the situation.

At the same time, organizations must ensure the privacy and security of the growing volumes of confidential information. Government and industry regulations from around the world, such as the Health Insurance Portability and Accountability Act (HIPAA), the Personal Information Protection and Electronic Documents Act (PIPEDA) and the Payment Card Industry Data Security Standard (PCI DSS) require organizations to protect personal information no matter where it lives-even in test and Development environments.

Archive data

Many organizations confuse about the difference between archiving data and backing up the data, Archiving preserves the data providing a long-term repository of information that can be used by litigation and audit teams. By contrast, backing up data means taking entire copy of current data to different place to prevent data from unfortunate deletion or disaster recovery and restoration of deleted files, Backups are often retained for a short time, until a fresh backup replaces the existing backup, Archiving complements backups by removing old, redundant and infrequently accessed data from a system and by reducing the size of databases and their backups.

Approximately 75 percent of the data stored is typically inactive, rarely accessed by any user, process or application. An estimated 90 percent of all data access requests are serviced by new data usually data that is less than a year old. With an effective archiving strategy, organizations can protect old data and comply with data retention rules while reducing costs and enhancing system performance. In an attempt to meet archiving needs, some organizations simply back up data to a Hadoop environment. But this kind of backup will not ensure that data will be fully protected or remain query-able, the way a true archive would. With an effective data lifecycle management solution, companies can create an archive that protects data, meets compliance standards, and supports queries and reporting. An emerging trend is for organizations to use Hadoop

As a lower-cost storage alternative for archives.

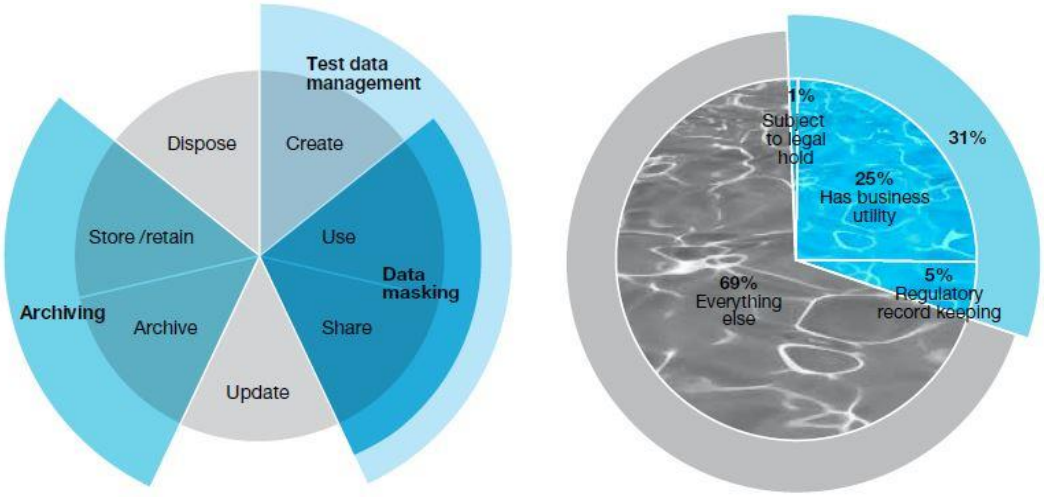


Figure 2.1: Big data life Cycle in Action

The data lifecycle having multiple phases including Create, Use, Share, Update, Archive, Store/ Retain, Dispose etc. where main there is outer 3 key player phases Test data management, data masking, Archiving.

Archiving: Policies are prepared to keep important data elements for reference and for future use while deleting data that is no longer necessary to support the legal needs of an organization.

It also prepares to improve performance of warehouse by archiving dormant data in a data warehouse with different databases.

Test data management: the data contains is might be newly created or historical data and might be real time data, it do testing based on the type of data.

Data masking: In big data, privacy and security must be implemented at all, some information like credit card / Debit card No, all this sensitive information should be masked using highly secure technique.

1.7 Big data Architecture

Data flow architecture

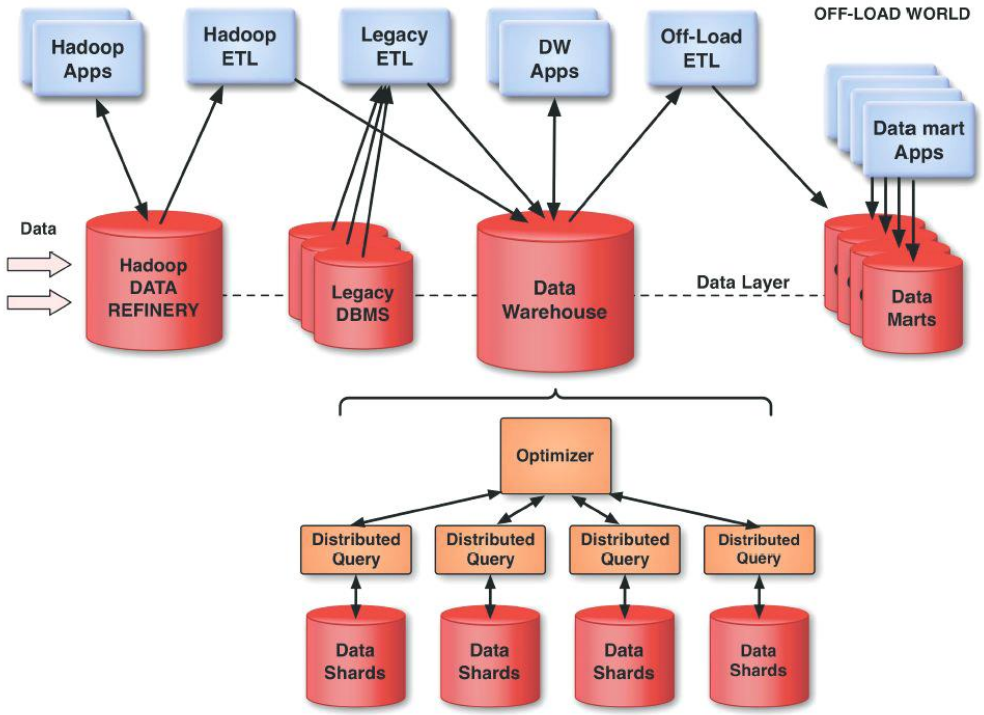


Figure 2.2**:** Data flow architecture, hadoop at data layer [2]

This data flow architecture explains that different format of data can come from Hadoop eco-system apps like Pig,Hive, Accumlo, sqoop, floom, oziee, spark,mahout, etc. This all apps produce different kinds of data at different times that data communicates with HADOOP DATA REFINERY.
Various other Hadoop ETL tools such as Talend Big data, Informatica Big data edition also takes data from DATA WAREHOUSE and communicates with HADOOP DATA REFINERY, In enterprise there are also legacy DBMS working with some other applications that might be target or source to the Legacy ETL tools.
DW Apps: data warehouse tools used to do OLAP operations such as Drill down, Drill up, Slice, Dice, and pivot with multidimensional cubes for Star and Snow flack schema. Popular Data warehouse tool including Business Intelligence are SAP Business Intelligence Objects, MicroStrategy Business Intelligence, Oracle Business Intelligence for Enterprise editions, Oracle Hyperion Business intelligence and financial Reports, IBM Cognos Business Intelligence etc.
This all BI DW applications are directly communicates with DATA WAREHOUSE.
Data can be transferred to data marts using traditional ETL Jobs and also with Hadoop ETL jobs and vice versa.
Distributed query language is used to do performance optimization of ETL Jobs with Shared data.

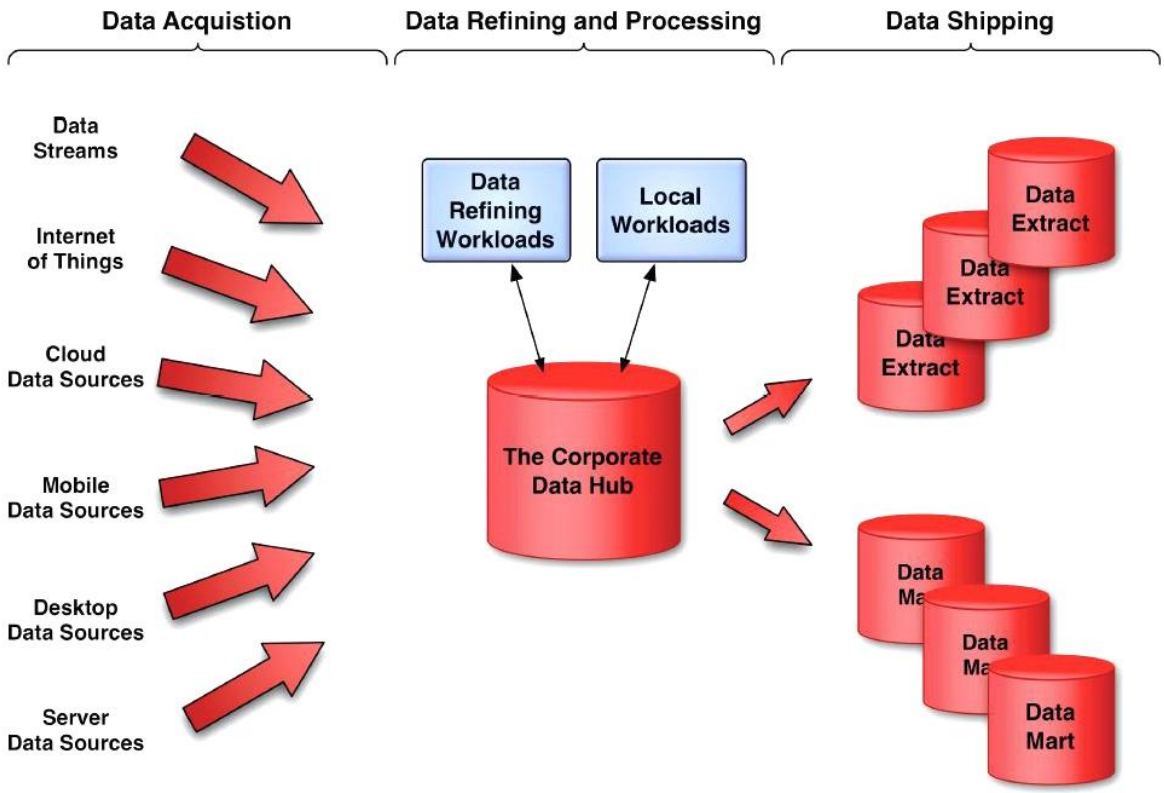


Figure 2.3: Big data Information Architecture [2]

For Data acquisition there are various sources such as Data Streams, Internet of Things, Cloud Data Source, Mobile Data Sources, Desktop Data Sources, Server Data sources etc, all this data sources produces different formats of data at different timing frequency that is part of Data Acquisition to consume different types of data for single data hub (THE CORPORATE DATA HUB).
All the BI and data analytics applications can run on the Data Hub. Additionally, all new Transactional applications can also run here, and should if feasible. It is for that reason, Incidentally, that the word "warehouse" is no longer appropriate for this large collection of Data. The applications that will not be able to use the Data Hub are as follows:

- Packaged software which is not able to use the capabilities and API of the Data Hub.
- Software that has inappropriate operational characteristics. For example, software that has inadequate network bandwidth or security characteristics to use the Data Hub as a data server.
- A good deal of office and personal software apps would also be inappropriate. Such software also has inappropriate operational characteristics, but it is worth mentioning separately. Note, however, that it is entirely possible that the Data Hub be used as a file sever to backup and serve the files used by such applications.

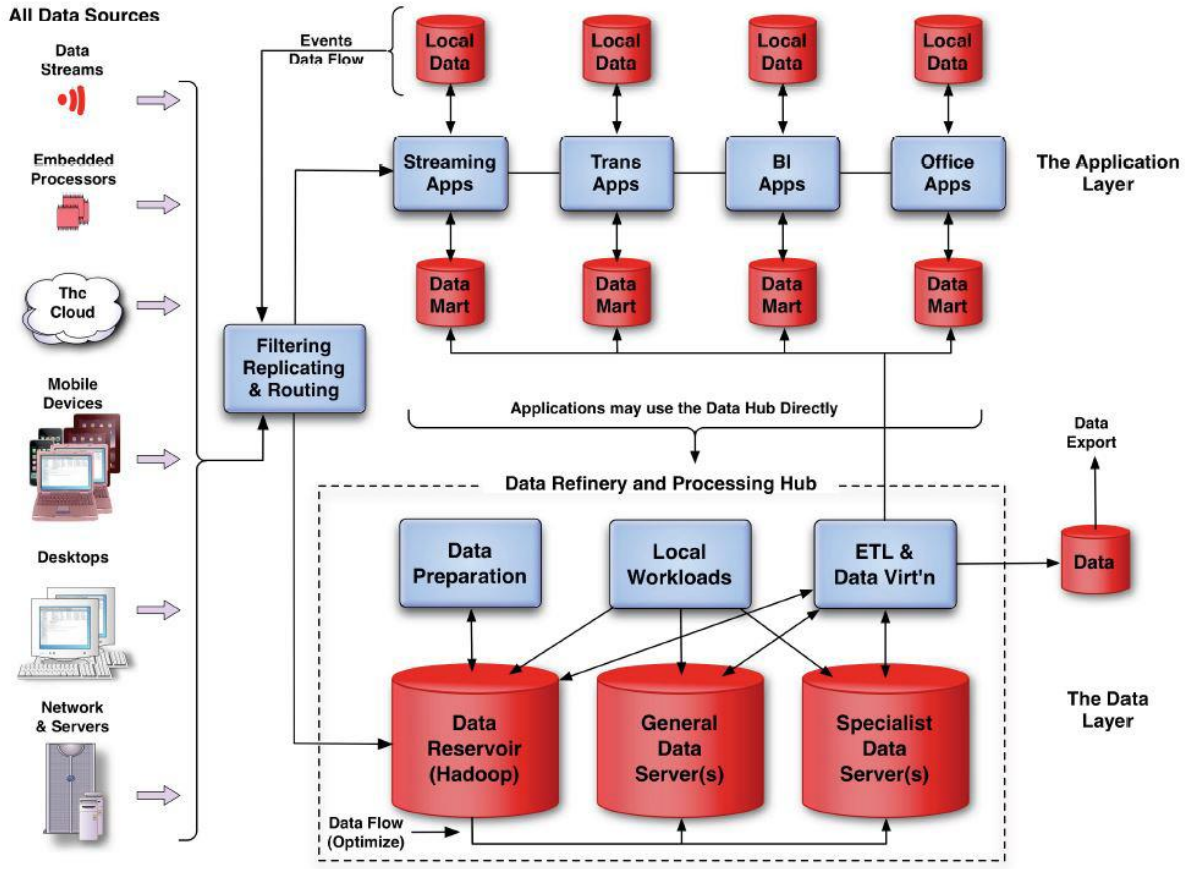


Figure 2.3: The data Refinery and processing hub in the data layer [2]

Big data architecture has various sources of unstructured data such as Data Streams, Sensors, Embedded processors, the cloud, mobile devices, Desktop, network & servers. All different sources have different kinds of data for pre-processing it after cleaning the data, after pre-process data can be used for Decision making and reporting.
Real time data is unpredictable so based on requirement we can store it in legacy RDBMS or if required than on Columnar NOSQL, Graph based NOSQL, or Key value pair NOSQL.
Real time data is Event based data, it occur based on Event(CEP[Complex Event processing] app), to consuming this types of data is complex in-terms of processing it. In Figure. All real time streaming data is passing out to Filtering, Replicating, Routing for further processing.
This process is the traffic cop for all newly emerging data. It knows where to direct the event, whether to duplicate it (for example, it could go both to a Streaming App and also be stored. within the Data Hub) or whether to simply delete it from the stream. The goal for this process is that it imposes almost no latency on the movement of data. This is necessary because some real-time streaming applications may have service levels that demand very very low latency.
In other words, this process needs to be fault tolerant and extremely fast.
In regards to applications in application layer either sends data request to data hub or use the data hub directly which becomes local workload, or they use data marts(data extracts) created and refreshed by ETL or data virtualization software, which access the data hub .
The same data virtualization capability can store and extract, export data from data hub.
The first point to note is that the Data Hub may have multiple engines. One useful way to think of this is that ingest into the Data Hub really is a refining process. First it is necessary to refine the data so that it is suitable to be processed.
And sometime we might be thinking some data being stored in Disk or SSD, some data may because high usage stored in direct memory immediately.
We might like to think of the Data Hub as a database, which it logically is, but physically it

will almost certainly consist of multiple engines. In our view, Hadoop (HDFS) and YARN will constitute one of those engines. HDFS will almost inevitably be the place where data is refined. Other data engines (one or more) will be optimized for specific workloads. It's worth noting that some of these workloads will involve analytic calculations as well as requests for data. Optimizing the performance of the Data Hub will thus be complex. It will involve optimization that is in a general sense unprecedented. We can think of it in the following way:

If data is available for processing but not stored optimally for its future use, and a query is received that touches that data, then the data will be read into memory.

However, it is very likely that it will be necessary to read the data into memory in order to refine the data. It is thus the case that with some data, the refining process is also the time at which the physical organization of the data should be determined. If it is not possible to do it then, it should be done the first time the data is touched.

In reality processing load in data hub is very complex, this illustrated in figure , there are some software application which may be capable of all this functions.

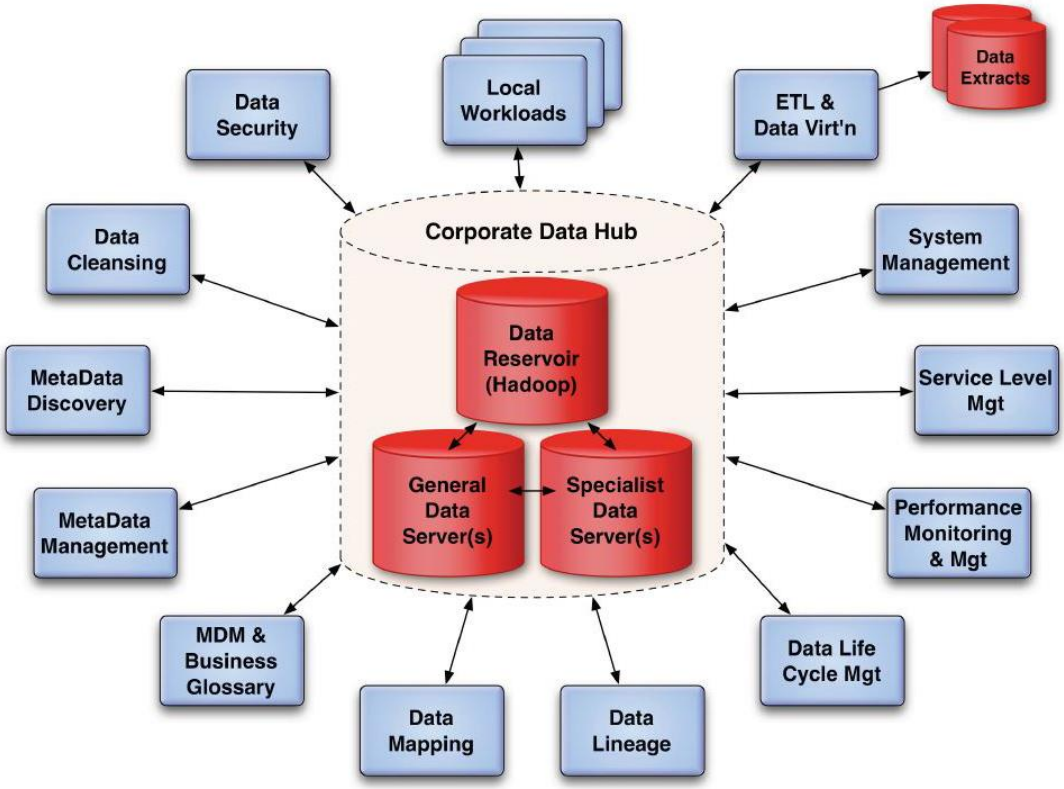


Figure 2.4 : Processes in Corporate Data hub. [2]

The various functions we have illustrated above are as follows:

Data Security: All data security procedures including encryption, both while data is at rest and while in motion, and all roles and responsibilities in respect to access rights need to be implemented from the moment that data enters the Corporate Data Hub.

Data Cleansing: Naturally, the full gamut of data cleansing activity from simple data Correction through data deduplication and disambiguation can and should be applied here.

Metadata Discovery: The data entering the Corporate Data Hub needs, before it is available for use, to have its metadata defined (to a given standard). This may involve a variety of processing activity from the application of standard patterns to known data sources to the use of semantics to determine, as accurately as possible, the meaning of the data. Some instances may require human intervention.

Metadata Management: The Corporate Data Hub will naturally accumulate a metadata resource (repository) that will need to be managed at the physical level and at the logical level. The activity here is primarily one of assembling metadata catalogs and/or taxonomies which can provide access to metadata both for software and for users.

MDM & Business Glossary: Master data management (MDM) is a natural extension of metadata management. It is collaboration on the business meaning of data and business terminology that may bring to light both terminology variances and data aliases. The goal is that data users (including software developers) can fully understand the data they have access to.
Data Mapping: Ideally, it will be possible to assemble and maintain a full data map of all data that is of interest to the corporation, not just the data stored within the Corporate Data Hub, but also all such data including metadata maps of data sources, data exports and data in motion.
Data Lineage: The provenance and lineage of all data needs to be captured and maintained. This is of particular importance for analytics activities since bad data can lead to wrong or inaccurate conclusions and actions.
Data Life Cycle Management: Given the above set of information it will become possible to proactively manage the life cycle of events and derived data to the point of data being retired and, if justified, deleted.
Performance Monitoring & Management: This can be thought of as the low-level management of the data engine(s) to optimize the performance of individual workloads and individual data engines.
Service Level Management: This is traditional service level management applied to the Corporate Data Hub. It involves the scheduling of workloads against available resources in order to meet agreed and targeted service levels.
System Management: This involves all other system management activities surrounding data flow, including fire-fighting, software management, IT asset management, network management and so on.
ETL & Data Virtualization: This is the export of data from the Data Hub both for apps within the corporate environment and for data customers elsewhere.

All but the last six of these functions concern data refinement. The critical ones are data Security, data cleansing and metadata discovery since these need to be applied before the data is usable. In some circumstances, when the metadata is very discoverable, a schema-on-read approach may be viable.
Metadata management, MDM and business glossary, data mapping and data lineage may be deferrable and may be part of a corporate effort to manage the data resource and improve its usability.
It could be said that the Big Data world of metadata is like the old world of metadata except that it has many more data sources – and just like data from packaged software, the business has no control over the data definitions of the new sources. We could even think of this as "big metadata." If Big Data means lots more data than we had before, then big metadata is lots more data sources.
In our view, data lineage is likely to become an increasingly important aspect of data management and particularly the management of the Data Hub, partly because of the increasing importance of data analytics, and partly because we are beginning to witness significant growth in the market for data itself. It is going to be progressively difficult to sell data at the best price without being able to declare its provenance and lineage.

1.8 What is Data warehousing and Data mining?

Data Warehousing:
Data warehousing provides architectures and tools for business executives to systematically organize, understand, and use their data to make strategic decision. Loosely, speaking a data warehouse refers to a database that is maintained separately from an organization's operational databases. Data warehouse systems allow for the integration of a variety of application systems. They support information processing by providing a solid platform of consolidated historical data for analysis.
According to William H. Inmon, a leading architect in the construction of data warehouse systems, "A data warehouse is a subject-oriented, integrated, time-variant, and nonvolatile collection of data in support of management's decision making process". This short, but comprehensive definition presents the major features of a data warehouse.

This short, but comprehensive definition presents the major features of a data warehouse. The four keywords, subject-oriented, integrated, time-variant, and non-volatile, distinguish data warehouses from other data repository systems, such as relational database systems,

transaction processing systems, and file systems. Let's take a closer look at each of these key features.

Subject-oriented: A data warehouse is organized around major subjects, such as customer, supplier, product and sales. Rather than concentrating on the day-to-day operations and transaction processing of an organization, a data warehouse focuses on the modeling and analysis of data for decision makers. Hence, data warehouse typically provide a simple and concise view around particular subject issues by excluding data that are not useful in the decision support process.
Integrated: A data warehouse is usually constructed by integrating multiple heterogeneous sources, such as relational databases, flat files, and on-line transaction records. Data cleaning and data integration techniques are applied to ensure consistency in naming conventions, encoding structures, attribute measures, and so on.
Time-variant: Data are stored to provide information from a historical perspective (e.g. past 5-10 years). Every key structure in the data warehouse contains, either implicitly or explicitly, an element of time.
Nonvolatile: A data warehouse is always a physically separate store of data transformed from the application data found in the operational environment. Due to this separation, a data warehouse does not require transaction processing, recovery, and concurrency control mechanisms. It usually requires only two operations in data accessing: initial loading of data and access of data.

Data Mining:
Almost all enterprises collect a variety of information electronically. In recent years there has been an explosive growth in the generation and storage of electronic information as more and more operations of enterprises are computerized due to technology advances in data acquisition and a continued decline in the data storage costs.
It has been reported that a large chain of stores like Wal-Mart in the USA with more than 4000 stores worldwide in 2004 has to deal with a database of perhaps more than 460 terabytes!
Most enterprises have started to realize that the information accumulated over the years constitutes important strategic asset. This intelligence can be the secret weapon on which the success of a business may depend. It is not a simple matter to discover business intelligence that might assist decision making. What is required are techniques that allow the enterprise to distil the most valuable information from mountains of accumulated data. The field of data mining provides such techniques.
Data mining or knowledge discovery in the databases (KDD) is a collection of exploration techniques based on advanced analytical methods and tools for handling a large amount of information.
Data mining may be defined as follows:
Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful and understandable patterns in large databases. The patterns must be actionable so that they may be used in an enterprise's decision making process.

If the amount of data is larger better chance of finding something novel and interesting, using data mining techniques. Data mining is an attempt at providing some of the intelligence that lets a manager understands his/her business better. OLTP is for day to day operations, could not be used for data mining. Data warehouse is not necessarily the best source for data mining analysis as the information required for access may not be available in the warehouse. Data mining is often a complex process and may require a variety of steps before some useful results are obtained. Often data pre-processing including data cleaning may be needed. In some cases, sampling of data and testing of various hypotheses may be required before data mining can start.

1.9 Data mining techniques

Data mining employs a number of techniques including the following:

- Association rules mining or market basket analysis
- Supervised classification
- Cluster analysis
- Web data mining

- Search engines
- Data warehousing and OLAP

1.10 Data warehouse Operations - Business Intelligence Tools

Following are OLAP operations on data cube.

1. **Roll-up:**
   Performs aggregation on a data cube by climbing-up a concept hierarchy for a dimension and dimension removal.
2. **Drill-down:**
   Navigates a data cube from less detailed data to more detailed data by stepping-down a concept hierarchy for a dimension introducing and additional dimensions.
3. **Slice:** performs a selection on one dimension of a data cube, resulting in a sub-cube.
4. **Dice:** performs a selection on multiple dimensions of a data cube.
5. **Pivot:** rotates the data axes in view also called rotate.

There are various Business Intelligence tools are available in market to accomplish Data Warehousing operations.
Following are popular Business Intelligence tools:

1. Oracle Business Intelligence for Enterprise Edition(OBIEE)
2. Oracle Hyperion Business Intelligence Suite.
3. SAP Business Intelligence Objects.
4. MicroStrategy Business Intelligence.
5. IBM Cognos suite.

Chapter 2
Growth of Data

2.1 Rapid growth of Unstructured data - Source of Big data

Today, growth of data is going on more fast, over here time is known as Internet time not human time. Every change done based on Time of Internet.
Rapid growth of unstructured data is much faster due to following key player components:

- Social Media
- Sensors
- E-commerce Websites
- Search Engine / Web crawler
- Email
- Instant Messaging
- Online Maps
- Air Traffic Controller(ATC)
- Data produces by Big Giant Companies
- Stock Exchange data
- Telco data
- Sports data
- Machine log data
- Documents on Cloud
- Health Care data
- IoT(Internet of Things)

Some facts based on Rapid growth of big data

- **Air Traffic Controller (ATC)** processes big amount of unstructured data, even "Four Engine jumbo jet currently generates around 540 Terabytes of data on single Atlantic crossing and the IoT has only just begun". [3]
- **Online Maps**
  - When you go to Google map using mobile, you are sending information to Google where are you? , how speed you are going?
  - I-Phone Collects data that where people are travelling (Aggregating together).
- **IoT(Internet of Things)** [Every device communicating each other] = Information Technology + Operations technology
  - The future will be data driven but who will drive the data?
    - **We're!**
  - Ex. Siri in Apple i-phone generates data.
  - Google glass generates data.
  - By 2014, it's anticipated there will be 420 Million wearable, wireless health monitors.
- **Social Media :**
  - **Facebook:**
    - Facebook stores, accesses and anlyzes 30+ Petabytes of user generated data.
    - Every month there are: [4]
      - 1.39 Billion on Facebook
      - 700 Million on Groups
      - 700 Million on WhatsApp
      - 500 Million on Messanger

        - 300 Million on Instagram

    - Every day there are:
        - 890 Million on Facebook
        - 1+ Billion Searches on Facebook
        - 3+ Billion Video views on Facebook
        - 7+ Billion likes on Facebook
        - 30+ Billion Messages on WhatsApp
  - **LinkedIn:** Processes and mines petabytes of user data to power "**People you may know**"
    - 2.1 million groups have been created.
  - **YouTube:** [**5**]
    - 4 Billion+ hours of video are watched on youtube each month.
    - Users upload 100 hours of new videos per minute.
    - Each month, more than 1 billion unique users access YouTube.
    - Over 6 billion hours of video are watched each month, which corresponds to almost an hour for every person on Earth. This figure is 50% higher than that generated in the previous year.
  - **Twitter:** [**6**]
    - 400 Million Tweets are sent per day by about 200 million monthly active users.
    - The site has over 645 million users.
    - The site generates 175 million tweets per day.
  - **Foursquare:** [**7**]
    - This site is used by 45 million people worldwide.
    - This site gets over 5 billion check-ins per day.
    - Every minute, 571 new websites are launched.
  - **Google+** [**8**]
    - 1 billion accounts have been created.
  - **Tumblr:**
    - Blog owners publish 27,000 new posts per minute.
  - **Instagram:**
    - Users share 40million photos per day.
  - **Flickr:**
    - Users upload 3,125 new photos per minute.
  - **WordPress:** [**9**]
    - Bloggers publish near 350 new blogs per minute.
- **Sensors:** Modern cars have close to 100 Sensors that monitors items such as fuel level and tire pressure, even who sits in car little boy, pregnant woman !
  - Each and every sensor produce huge amount of data that is purely unstructured data.
- **E-commerce Websites:**
  - E-Commerce website produces huge amount of data by produces user log for prediction with Click Stream analysis.
  - **Amazon:** Crunches click-stream and historical user data to recommended products.
  - **Akamai Analyzes:** 75 million events per day to better target advertisements.
  - **JP Morgan Chase & Co.:** Analyzes web logs, transaction data, and social media to detect fraudulent activity.
- **Search Engine / Web crawler** consumes huge amount of unstructured data by indexing and crawling the web pages and resources.
- **Spamming Email** can also produce huge amounts of unstructured unvalued data.
- **Instant Messaging apps** are producing huge amounts of data by containing text, video, images, location maps etc.

- **Telco data:** Every telecommunication companies track user‘s usage to give best offers.
- **Sports data:** Every event stored huge unstructured data by each and every player’s position and related score using various filtration methods. Such as Fifa World cup etc.
- **Machine log data:** every satellite produces a huge amount of unstructured data to show information such as Weather data.
- **Documents of Cloud:** Currently cloud contains huge amount of documents, spreadsheet and other file formats.
- **Stock Exchange data:** New York Stock Exchange produces 1 TB data per day. [37]
- **Health Care:** Treato taps big data to help researchers and physicians better determined patient treatments.

Other:

- 40 Zettabytes of data will be created by 2020, an increase of 300 times form 2005
- 6 billion people have cell phones , world population 7 Billion
- Most companies in the US have at least 100 terabytes of data stored.
- It’s estimated that 2.5 quintillion bytes of data are created each day.
- By 2016, its projected there will be 18.9 billion network connections almost 2.5 connections per person on earth.
- The New York Times processes 4TB worth of raw images into 11 Million finished pdfs in 24 hours

Chapter 3
Tools and technologies used for Big data

3.1 Hadoop

Introduction: [11]
The Apache Hadoop project develops open-source software for reliable, scalable, distributed computing.

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

The project includes these modules:

Hadoop Common: The common utilities that support the other Hadoop modules.
Hadoop Distributed File System (HDFS): A distributed file system that provides high-throughput access to application data.
Hadoop YARN: A framework for job scheduling and cluster resource management.
Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.
Other Hadoop-related projects at Apache include:

Ambari: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters which includes support for Hadoop HDFS, Hadoop MapReduce, Hive, HCatalog, HBase, ZooKeeper, Oozie, Pig and Sqoop. Ambari also provides a dashboard for viewing cluster health such as heatmaps and ability to view MapReduce, Pig and Hive applications visually alongwith features to diagnose their performance characteristics in a user-friendly manner.
Avro: A data serialization system.
Cassandra: A scalable multi-master database with no single points of failure.
Chukwa: A data collection system for managing large distributed systems.
HBase: A scalable, distributed database that supports structured data storage for large tables.
Hive: A data warehouse infrastructure that provides data summarization and ad hoc querying.
Mahout: A Scalable machine learning and data mining library.
Pig: A high-level data-flow language and execution framework for parallel computation.
Spark: A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
Tez: A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases. Tez is being adopted by Hive™, Pig™ and other frameworks in the Hadoop ecosystem, and also by other commercial software (e.g. ETL tools), to replace Hadoop™ MapReduce as the underlying execution engine.
ZooKeeper: A high-performance coordination service for distributed applications.

The Hadoop framework itself is mostly written in the Java programming language, with some native code in C and command line utilities written as shell-scripts. For end-users, though MapReduce Java code is common, any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of the user's program. Other related projects expose other higher level user interfaces.

Prominent corporate users of Hadoop include Facebook and Yahoo. It can be deployed in traditional onsite datacenters as well as via the cloud; e.g., it is available on Microsoft Azure, Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3), Google App Engine and IBM Bluemix cloud services.

Architecture:

Hadoop consists of the Hadoop Common package, which provides filesystem and OS level abstractions, a MapReduce engine (either MapReduce/MR1 or YARN/MR2)[16] and the Hadoop Distributed File System (HDFS). The Hadoop Common package contains the necessary Java ARchive (JAR) files and scripts needed to start Hadoop. The package also provides source code, documentation, and a contribution section that includes projects from the Hadoop Community.

For effective scheduling of work, every Hadoop-compatible file system should provide location awareness: the name of the rack (more precisely, of the network switch) where a worker node is. Hadoop applications can use this information to run work on the node where the data is, and, failing that, on the same rack/switch, reducing backbone traffic. HDFS uses this method when replicating data to try to keep different copies of the data on different racks. The goal is to reduce the impact of a rack power outage or switch failure, so that even if these events occur, the data may still be readable.

Hadoop cluster
A multi-node Hadoop cluster
A small Hadoop cluster includes a single master and multiple worker nodes. The master node consists of a JobTracker, TaskTracker, NameNode and DataNode. A slave or worker node acts as both a DataNode and TaskTracker, though it is possible to have data-only worker nodes and compute-only worker nodes. These are normally used only in nonstandard applications. Hadoop requires Java Runtime Environment (JRE) 1.6 or higher. The standard startup and shutdown scripts require that Secure Shell (ssh) be set up between nodes in the cluster.

In a larger cluster, the HDFS is managed through a dedicated NameNode server to host the file system index, and a secondary NameNode that can generate snapshots of the namenode's memory structures, thus preventing file-system corruption and reducing loss of data. Similarly, a standalone JobTracker server can manage job scheduling. In clusters where the Hadoop MapReduce engine is deployed against an alternate file system, the NameNode, secondary NameNode, and DataNode architecture of HDFS are replaced by the file-system-specific equivalents.

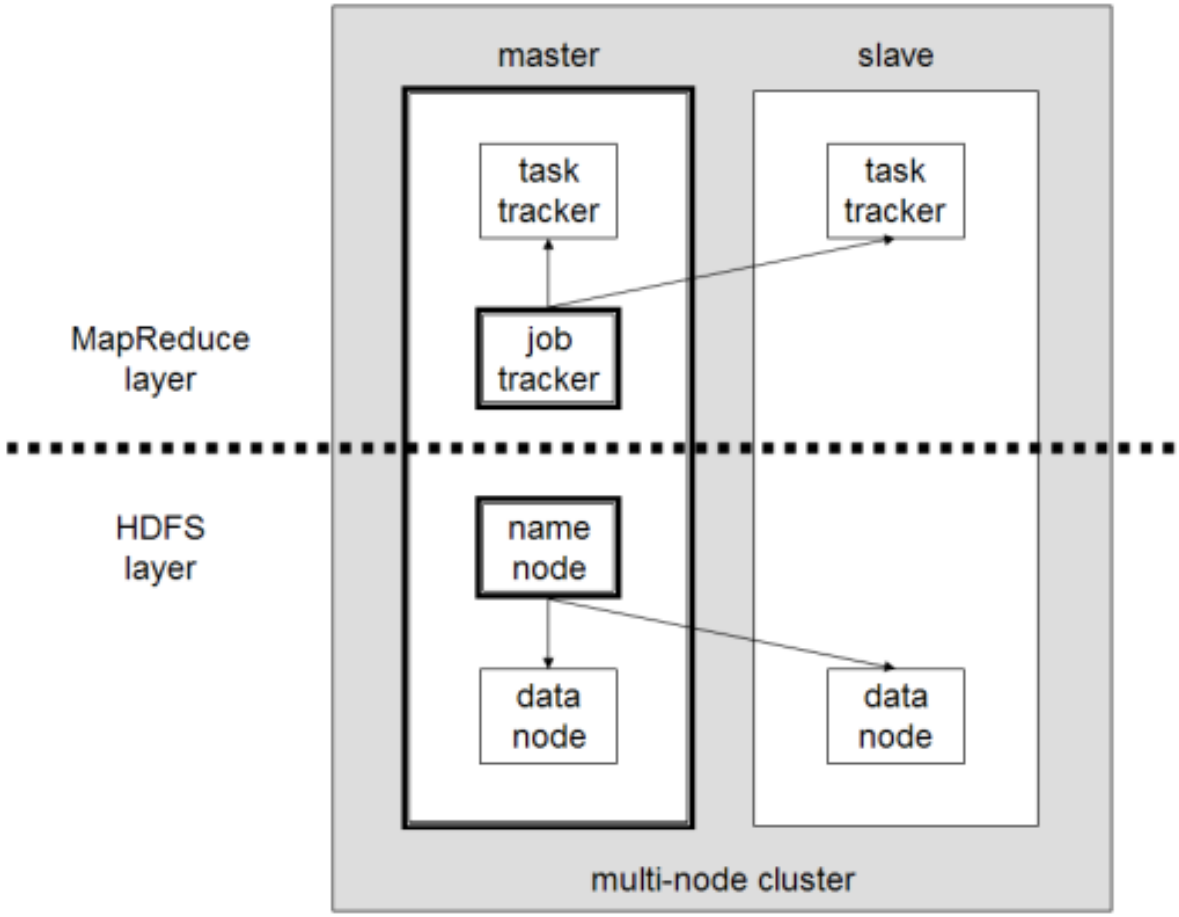


Figure 3.1: Multi node hadoop Cluster

Hadoop on the cloud:
Hadoop can be deployed in a traditional onsite datacenter as well as in the cloud. The cloud allows organizations to deploy Hadoop without hardware to acquire or specific setup expertise. Vendors who currently have an offer for the cloud include Microsoft, Amazon, and Google.

Hadoop on Microsoft Azure
Azure HDInsight is a service that deploys Hadoop on Microsoft Azure. HDInsight uses a Windows-based Hadoop distribution that was jointly developed with Hortonworks and allows programming extensions with .NET (in addition to Java). By deploying HDInsight in the cloud, organizations can spin up the number of nodes they want and only get charged for the compute and storage that is used. Hortonworks implementations can also move data from the on-premises datacenter to the cloud for backup, development/test, and bursting scenarios.

Hadoop on Amazon EC2/S3 services
It is possible to run Hadoop on Amazon Elastic Compute Cloud (EC2) and Amazon Simple Storage Service (S3). As an example The New York Times used 100 Amazon EC2 instances and a Hadoop application to process 4 TB of raw image TIFF data (stored in S3) into 11 million finished PDFs in the space of 24 hours at a computation cost of about $240 (not including bandwidth).

There is support for the S3 file system in Hadoop distributions, and the Hadoop team generates EC2 machine images after every release. From a pure performance perspective, Hadoop on S3/EC2 is inefficient, as the S3 file system is remote and delays returning from every write operation until the data is guaranteed not lost. This removes the locality advantages of Hadoop, which schedules work near data to save on network load.

Amazon Elastic MapReduce

Elastic MapReduce (EMR) was introduced by Amazon in April 2009. Provisioning of the Hadoop cluster, running and terminating jobs, and handling data transfer between EC2(VM) and S3(Object Storage) are automated by Elastic MapReduce. Apache Hive, which is built on top of Hadoop for providing data warehouse services, is also offered in Elastic MapReduce.

MapReduce [13]

MapReduce is a programming model and an associated implementation for processing and generating large data sets with a parallel, distributed algorithm on a cluster.

A MapReduce program is composed of a Map() procedure that performs filtering and sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() procedure that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies). The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance.

The model is inspired by the map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as in their original forms. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine once. As such, a single-threaded implementation of MapReduce (such as MongoDB) will usually not be faster than a traditional (non-MapReduce) implementation, any gains are usually only seen with multi-threaded implementations. Only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance features of the MapReduce framework come into play, is the use of this model beneficial. Optimizing the communication cost is essential to a good MapReduce algorithm.

MapReduce libraries have been written in many programming languages, with different levels of optimization. A popular open-source implementation that has support for distributed shuffles is part of Apache Hadoop. The name MapReduce originally referred to the proprietary Google technology, but has since been genericized.

Who uses Hadoop ?[14]

Nowadays, Every big giant companies uses Hadoop from Telco to Government also. But some popular top ranking hadoop vendors and users mentioned over here.

Hadoop Vendors

- Apache Hadoop
- Cloudera
- Hortonworks
- MAPR Technologies
- EMC^2
- IBM
- Amazon Web Service
- Intel

IT Vendors

- Teradata – Unified data architecture
- Oracle – Big data appliances X3 – 2
- IBM Big Insights
- HP Vertica
- EMC Pivotal Greenplum

Hadoop + IT Vendors

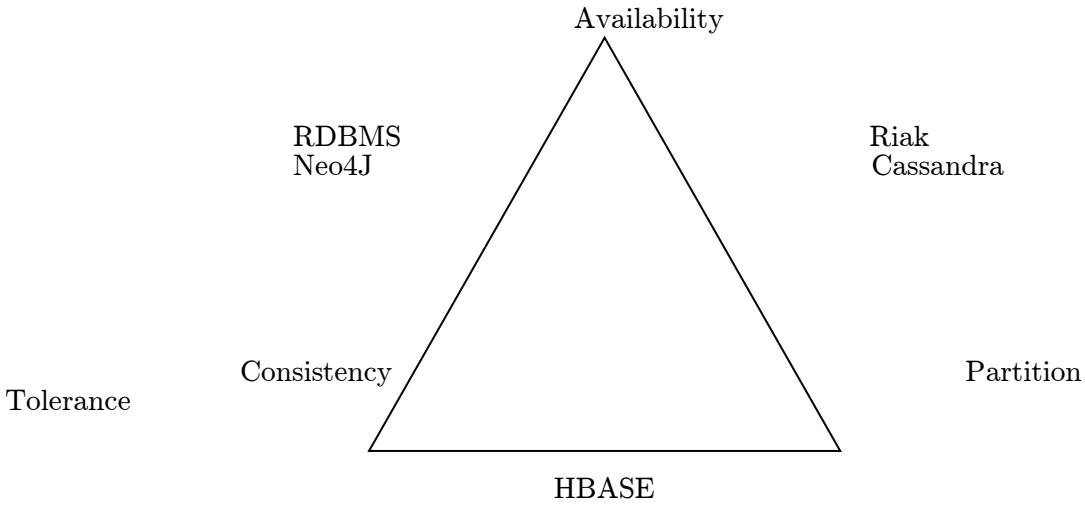
- Zettaset
- Datameer

3.2 Introduction to NOSQL

NOSQL does not mean NO SQL, but it means NOT ONLY SQL, it's a schema less database or NON Relational database.
A NoSQL (often interpreted as Not only SQL) database provides a mechanism for storage and retrieval of data that is modeled in means other than the tabular relations used in relational databases. Motivations for this approach include simplicity of design, horizontal scaling, and finer control over availability. The data structures used by NoSQL databases (e.g. key-value, graph, or document) differ from those used in relational databases, making some operations faster in NoSQL and others faster in relational databases. The particular suitability of a given NoSQL database depends on the problem it must solve.

NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also called "Not only SQL" to emphasize that they may also support SQL-like query languages. Many NoSQL stores compromise consistency (in the sense of the CAP theorem) in favor of availability and partition tolerance. Barriers to the greater adoption of NoSQL stores include the use of low-level query languages, the lack of standardized interfaces, and huge investments in existing SQL. Most NoSQL stores lack true ACID transactions, although a few recent systems, such as FairCom c-treeACE, Google Spanner (though technically a NewSQL database), FoundationDB and OrientDB have made them central to their designs.

CAP Theorem:
It's impossible for a distributed computer system to simultaneously provide all three of the following guarantees.



"No Schema", "No transaction", "No Language" , A reboot of data systems focusing on just high – throughput reads and writes. But a clear trend towards re-introducing schemas, languages, transaction at full scale. Ex. Google Spanner system.

Types of NoSQL databases [15]
There have been various approaches to classify NoSQL databases, each with different categories and subcategories. Because of the variety of approaches and overlaps it is difficult to get and maintain an overview of non-relational databases. Nevertheless, a basic classification is based on data model. A few examples in each category are:

Column: Accumulo, Cassandra, Druid, HBase, Vertica
Document: Clusterpoint, Apache CouchDB, Couchbase, MarkLogic, MongoDB, OrientDB
Key-value: CouchDB, Dynamo, FoundationDB, MemcacheDB, Redis, Riak, FairCom c-treeACE, Aerospike, OrientDB, MUMPS
Graph: Allegro, Neo4J, InfiniteGraph, OrientDB, Virtuoso, Stardog
Multi-model: OrientDB, FoundationDB, ArangoDB, Alchemy Database, CortexDB

Document store
Nested values, extensible records(Think XML,YAML, JSON,BSON)
Key-Value
Associate array of a set of key-vlaue pairs, No schema, no exposed nesting.
Graph
This kind of database is designed for data whose relations are well represented as a graph (elements interconnected with an undetermined number of relations between them). The kind of data could be social relations, public transport links, road maps or network topologies, for example.
Multi-model database
Most database management systems are organized around a single data model that determines how data can be organized, stored, and manipulated. In contrast, a multi-model database is designed to support multiple data models against a single, integrated backend.[1] Document, graph, relational, and key-value models are examples of data models that may be supported by a multi-model database.

Column
A column of a distributed data store is a NoSQL object of the lowest level in a keyspace. It is a tuple (a key-value pair) consisting of three elements:[1]

Unique name: Used to reference the column
Value: The content of the column. It can have different types, like AsciiType, LongType, TimeUUIDType, UTF8Type among others.
Timestamp: The system timestamp used to determine the valid content.

NOSQL Myths
- NOSQL Databases are read only.
- NOSQL Databases are a part of hadoop and always with hadoop.
- NOSQL does not have any schema at all.

Advantages
- Ability to horizontally scale "Simple operation" throughput over many servers.
- Simple key lookups read/write of 1 or few records.
- The Ability to replicate and partition data over many servers.
- Consider "sharing" and "Horizontal" partitioning to be synonyms.
- A Simple API no query language
- A weaker concurrency model than ACID transactions.
- Efficient use of distributed indexes and RAM for data storage.
- The ability to dynamically add new attributes to data records.
    - i.e. – Scalable
    - NOSQL
    - No transaction
    - No Schema

Disadvantages
- Can not handle complex joins.
- No Support for complex transactions.
- Constraint support is not there.

Future:
Is it the end of RDBMS's ?
No ! We need both!

3.3 DSMS(Data Stream Management System)

A Data stream management system (DSMS) is a computer program to manage continuous data streams. It is similar to a database management system (DBMS), which is, however, designed for static data in conventional databases. A DSMS also offers a flexible query processing so that the information need can be expressed using queries. However, in contrast to a DBMS, a DSMS executes a continuous query that is not only performed once, but is permanently installed. Therefore, the query is continuously executed until it is explicitly

uninstalled. Since most DSMS are data-driven, a continuous query produces new results as long as new data arrive at the system. This basic concept is similar to Complex event processing so that both technologies are partially coalescing.

One of the most important features of a DSMS is the possibility to handle potentially infinite and rapidly changing data streams by offering a flexible processing at the same time, although there are only limited resources like a limited main memory. The following table provides various principles of DSMS and compares them to traditional DBMS. [17]

Database management system (DBMS)	Data stream management system (DSMS)
Persistent data (relations)	volatile data streams
Random access	Sequential access
One-time queries	Continuous queries
(theoretically) unlimited secondary storage	limited main memory
Only the current state is relevant	Consideration of the order of the input
relatively low update rate	potentially extremely high update rate
Little or no time requirements	Real-time requirements
Assumes exact data	Assumes outdated/inaccurate data
Plannable query processing	Variable data arrival and data characteristics

3.4 Visualization Tools & Technologies

Without visualizing data to take interactive decision from dataset is unpredictable, to present reports and final decision data visualization tools can help to easily explain the interactive report in meeting, Data visualizations are everywhere today. From creating a visual representation of data points to impress potential investors, report on progress, or even visualize concepts for customer segments, data visualizations are a valuable tool in a variety of settings. When it comes to big data, weak tools with basic features don't cut it.

Some popular tools are as below:

- Polymaps
- Nodebox
- Flot
- Processing
- Tabeulo
- Tangle
- D3
- FF Chartwell
- SAS Visual Analytics
- Raphael
- Linkscape
- visual.ly
- Revisit
- Google fusion tables
- Dipity
- Many eyes
- WkiMindMap
- HTML Graph
- Axiis
- Tweet Spectrum
- Wordle
- Tag Crowd
- Vuvox
- yoooouuutuuube
- Anaytics Visualization

- Newsmap
- Linkedin labs- InMaps
- Wolfram Alpha
- Google Public Data Explorer

3.5 ETL Tools

ETL tools are used to do pre-processing stuff like data cleaning, data quality, data integration, data migration, data filtering etc. The duty of any ETL tools is to Extract, Transform, load unstructured / structured data.

Following are list of popular ETL tools:

Open Source ETL tools:

1. Apatar
2. CloverETL
3. GeoKettle
4. Jaspersoft ETL
5. KETL
6. Pentaho's Data Integration
7. Talend

Commercial ETL Tools:

1. IBM (Information Server Infosphere platform)

   Advantages:

   - strongest vision on the market, flexibility
   - progress towards common metadata platform
   - high level of satisfaction from clients and a variety of initiatives

   Disadvantages:

   - difficult learning curve
   - long implementation cycles
   - became very heavy (lots of GBs) with version 8.x and requires a lot of processing power

2. Informatica PowerCenter

   Advantages:

   - most substantial size and resources on the market of data integration tools vendors
   - consistent track record, solid technology, straightforward learning curve, ability to address real-time data integration schemes
   - Informatica is highly specialized in ETL and Data Integration and focuses on those topics, not on BI as a whole
   - focus on B2B data exchange

   Disadvantages:

   - several partnerships diminishing the value of technologies
   - Limited experience in the field.

3. Microsoft (SQL Server Integration Services)

Advantages:

- broad documentation and support, best practices to data warehouses
- ease and speed of implementation
- standardized data integration
- real-time, message-based capabilities
- relatively low cost - excellent support and distribution model

Disadvantages:

- Problems in non-Windows environments. Takes over all Microsoft Windows limitations.
- unclear vision and strategy

4. Oracle (OWB and ODI)

Advantages:

- based on Oracle Warehouse Builder and Oracle Data Integrator – two very powerful tools;
- tight connection to all Oracle datawarehousing applications;
- Tendency to integrate all tools into one application and one environment.

Disadvantages:

- focus on ETL solutions, rather than in an open context of data management;
- tools are used mostly for batch-oriented work, transformation rather than real-time processes or federation data delivery;
- long-awaited bond between OWB and ODI brought only promises - customers confused in the functionality area and the future is uncertain

5. SAP BusinessObjects (Data Integrator / Data Services)

Advantages:

- integration with SAP
- SAP Business Objects created a firm company determined to stir the market;
- Good data modeling and data-management support;
- SAP Business Objects provides tools for data mining and quality; profiling due to many acquisitions of other companies.
- Quick learning curve and ease of use

Disadvantages:

- SAP Business Objects is seen as two different companies
- Uncertain future. Controversy over deciding which method of delivering data integration to use (SAP BW or BODI).
- Business Objects Data Integrator (Data Services) may not be seen as a stand-alone capable application to some organizations.

6. SAS

Advantages:

- experienced company, great support and most of all very powerful data integration tool with lots of multi-management features

- can work on many operating systems and gather data through number of sources – very flexible
- great support for the business-class companies as well for those medium and minor ones

Disadvantages:

- misplaced sales force, company is not well recognized
- SAS has to extend influences to reach non-BI community
- Costly

7. Sun Microsystems

Advantages:

- Data integration tools are a part of huge Java Composite Application Platform Suite - very flexible with ongoing development of the products
- 'Single-view' services draw together data from variety of sources; small set of vendors with a strong vision

Disadvantages:

- relative weakness in bulk data movement
- limited mindshare in the market
- support and services rated below adequate

8. Sybase

Advantages:

- assembled a range of capabilities to be able to address a multitude of data delivery styles
- size and global presence of Sybase create opportunities in the market
- pragmatic near-term strategy - better of current market demand
- broad partnerships with other data quality and data integration tools vendors

Disadvantages:

- falls behind market leaders and large vendors
- gaps in many aspects of data management

9. Syncsort

Advantages:

- functionality; well-known brand on the market (40 years experience); loyal customer and experience base;
- easy implementation, strong performance, targeted functionality and lower costs

Disadvantages:

- struggle with gaining mind share in the market
- lack of support for other than ETL delivery styles
- unsatisfactory with lack of capability of professional services

10. Tibco Software

Advantages:

- message-oriented application integration; capabilities based on common SOA structures;
- support for federated views; easy implementation, support and performance

Disadvantages:

- scarce references from customers; not widely enough recognized for data integration competencies
- Lacking in data quality capabilities.

11. ETI

Advantages:

- proven and mature code-generating architecture
- one of the earliest vendors on the data integration market; support for SOA service-oriented deployments;
- successfully deals with large data volumes and a high degree of complexity, extension of the range of data platforms and data sources;
- customers' positive responses to ETI technology

Disadvantages:

- relatively slow growth of customer base
- Rather not attractive and inventive technology.

12. iWay Software

Advantages:

- offers physical data movement and delivery; support of wide range of adapters and access to numerous sources;
- well integrated, standard tools;
- reasonable ease of implementation effort

Disadvantages:

- gaps in specific capabilities
- relatively costly - not competitive versus market leaders

13. Pervasive Software

Advantages:

- many customers, years of experience, solid applications and support;
- good use of metadata
- Upgrade from older versions into newer is straightforward.

Disadvantages:

- inconsistency in defining the target for their applications;
- no federation capability;
- limitated presence due to poor marketing.

14. Open Text

Advantages

- Simplicity of use in less-structured sources
- Easy licensing for business solutions
- cooperates with a wide range of sources and targets
- increasingly high functionality

Disadvantages:

- limited federation, replication and data quality support; rare upgrades due to its simplicity;
- Weak real-time support due to use third party solutions and other database utilities.

15. Pitney Bowes Software

Advantages:

- Data Flow concentrates on data integrity and quality;
- supports mainly ETL patterns; can be used for other purposes too;
- Ease of use, fast implementation, and specific ETL functionality.

Disadvantages:

- Rare competition with other major companies, repeated rebranding trigger suspicions among customers.
- Narrow vision of possibilities even though Data Flow comes with variety of applications.
- Weak support, inexperienced service.

Chapter 4
Methods & Techniques in Big data

4.1 Batch processing

Batch data processing is an efficient way of processing high volumes of data is where a group of transactions is collected over a period of time. Data is collected, entered, processed and then the batch results are produced (Hadoop is focused on batch data processing). Batch processing requires separate programs for input, process and output. An example is payroll and billing systems.

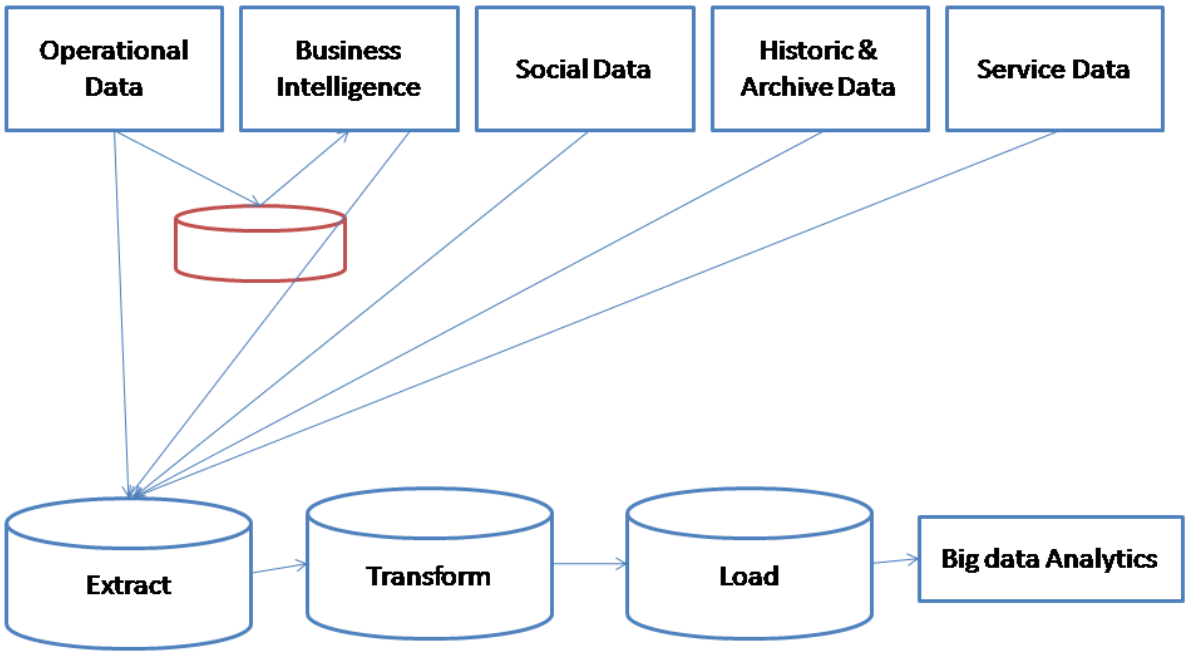


Figure 4.1 :Big data batch processing

In contrast, real time data processing involves a continual input, process and output of data. Data must be processed in a small time period (or near real time). Radar systems, customer services and bank ATMs are examples.

While most organizations use batch data processing, sometimes an organization needs real time data processing. Real time data processing and analytics allows an organization the ability to take immediate action for those times when acting within seconds or minutes is significant. The goal is to obtain the insight required to act prudently at the right time - which increasingly means immediately.
Complex event processing (CEP) combines data from multiple sources to detect patterns and attempt to identify either opportunities or threats. The goal is to identify significant events and respond fast. Sales leads, orders or customer service calls are examples.

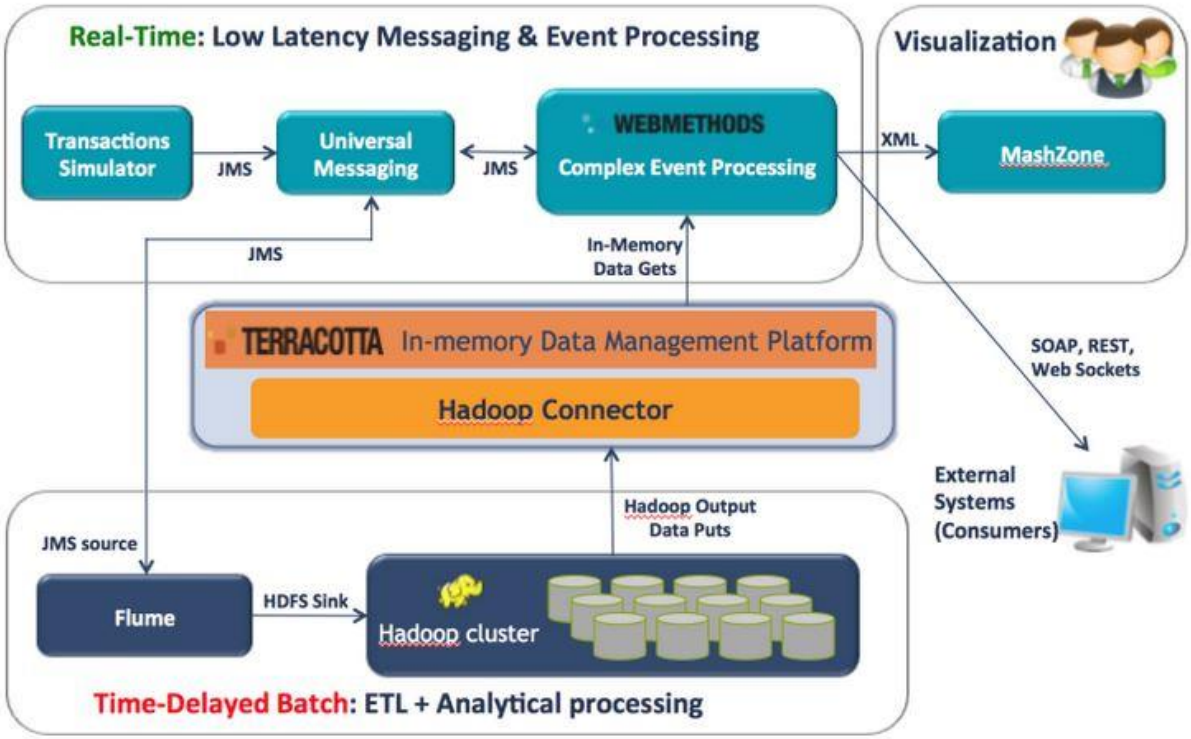


Figure 4.2 : ETL batch Processing in Hadoop Eco System [17]

In a Hadoop environment, the trick to providing near real time analysis is a scalable in-memory layer between Hadoop and CEP. Storm is an open source distributed real time computation system that processes streams of data. Storm can help with real time analytics, online machine learning, continuous computation, distributed RPC and ETL. Hadoop MapReduce processes "jobs" in batch while Storm processes streams in near real time. The idea is to reconcile real time and batch processing when dealing with large data sets. An example is detecting transaction fraud in near real time while incorporating data from the data warehouse or hadoop clusters.

Following are batch processing examples:

Solution	Developer	Type	Description
Storm	Twitter	Streaming	Twitter's new streaming big-data analytics solution
S4	Yahoo!	Streaming	Distributed stream computing platform from Yahoo!
Hadoop	Apache	Batch	First open source implementation of the MapReduce paradigm
Spark	UC Berkeley AMPLab	Batch	Recent analytics platform that supports in-memory data sets and resiliency
Disco	Nokia	Batch	Nokia's distributed MapReduce framework
HPCC	LexisNexis	Batch	HPC cluster for big data

Figure 4.3 : batch processing projects

4.2 Interactive analytics & Stream processing

"Streaming processing" is the ideal platform to process data streams or sensor data (usually a high ratio of event throughput versus numbers of queries), whereas "complex event processing" (CEP) utilizes event-by-event processing and aggregation (e.g. on potentially out-of-order events from a variety of sources – often with large numbers of rules or business logic). CEP engines are optimized to process discreet "business events" for example, to compare out-of-order or out-of-stream events, applying decisions and reactions to event patterns, and so on. For this reason multiple types of event processing have evolved, described as queries, rules and procedural approaches (to event pattern detection). The focus of this article is on stream processing.
Stream processing is designed to analyze and act on real-time streaming data, using "continuous queries" (i.e. SQL-type queries that operate over time and buffer windows). Essential to stream processing is Streaming Analytics, or the ability to continuously calculate mathematical or statistical analytics on the fly within the stream. Stream processing solutions are designed to handle high volume in real time with a scalable, highly available and fault tolerant architecture. This enables analysis of data in motion.

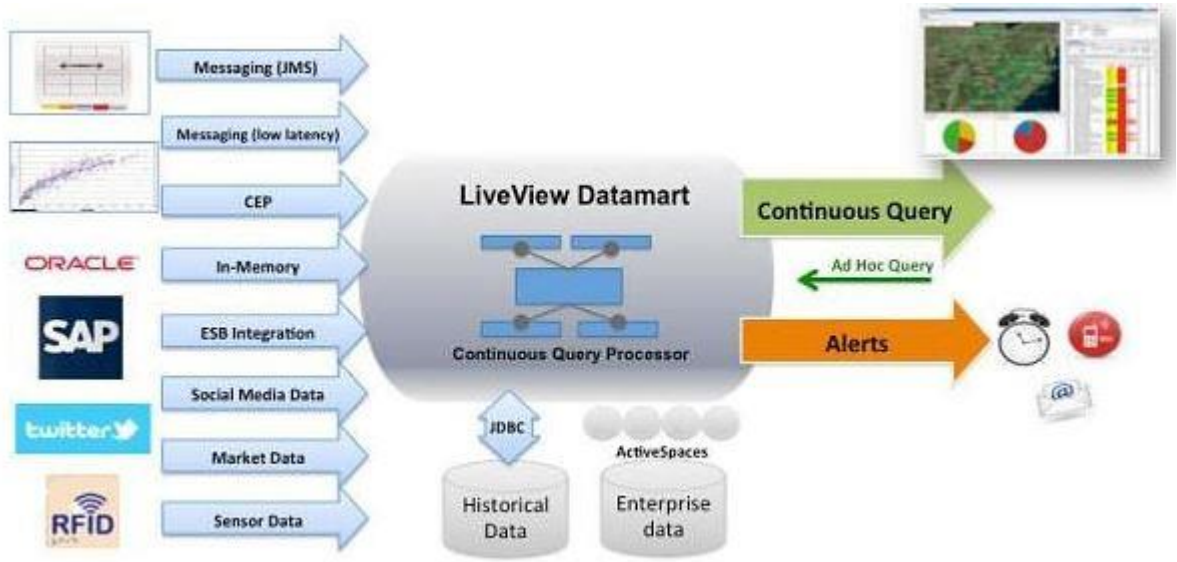


Figure 4.4 : Stream Processing Architecture

In contrast to the traditional database model where data is first stored and indexed and then subsequently processed by queries, stream processing takes the inbound data while it is in flight, as it streams through the server. Stream processing also connects to external data sources, enabling applications to incorporate selected data into the application flow, or to update an external database with processed information.

The rapid growth in data volumes requires new computer systems that scale out across hundreds of machines. While early frameworks, such as MapReduce, handled large-scale batch processing, the demands on these systems have also grown. Users quickly needed to run (1) more interactive ad-hoc queries, (2) more complex multi-pass algorithms (e.g. machine learning and graph processing), and (3) real-time processing on large data streams.

A recent development in the stream processing industry is the invention of the "live data mart" which provides end-user, ad-hoc continuous query access to this streaming data that's aggregated in memory. Business user-oriented analytics tools access the data mart for a continuously live view of streaming data. A live analytics front ends slices, dices, and aggregates data dynamically in response to business users' actions, and all in real time.

A stream processing solution has to solve different challenges:

- Processing massive amounts of streaming events (filter, aggregate, rule, automate, predict, act, monitor, alert)
- Real-time responsiveness to changing market conditions
- Performance and scalability as data volumes increase in size and complexity
- Rapid integration with existing infrastructure and data sources: Input (e.g. market data, user inputs, files, history data from a DWH) and output (e.g. trades, email alerts, dashboards, automated reactions)
- Fast time-to-market for application development and deployment due to quickly changing landscape and requirements
- Developer productivity throughout all stages of the application development lifecycle by offering good tool support and agile development
- Analytics: Live data discovery and monitoring, continuous query processing, automated alerts and reactions
- Community (component / connector exchange, education / discussion, training / certification)
- End-user ad-hoc continuous query access
- Alerting
- Push-based visualization

Tools for Stream analysis

1. Apache storm
2. Apache Spark
3. IBM InfoSphere Streams
4. TIBCO StreamBase
5. Apache Samza
6. Esper
7. AWS Kinesis
8. DataTorrent
9. Apama
10. Oracle CEP
11. Sybase CEP

4.3 Sentiment Analytics

Sentiment analysis refers to various methods of examining and processing data in order to identify a subjective response, usually a general mood or a group's opinions about a specific topic. For example, sentiment analysis can be used to gauge the overall positivity toward a blog or a document, or to capture constituent attitudes toward a political candidate. Sentiment data is often derived from social media services and similar user-generated content, such as reviews, comments, and discussion groups. The data sets thus tend to grow large enough to be considered "big data."[18]
Today's consumers are heavily involved in social media, with users having accounts on multiple social media services. Social media gives users a platform to communicate effectively with friends, family, and colleagues, and also gives them a platform to talk about their favorite (and least favorite brands). This "unstructured" conversation can give businesses valuable insight into how consumers perceive their brand, and allow them to actively make business decisions to maintain their image.

Historically, unstructured data has been very difficult to analyze using traditional data warehousing technologies. New cost effective solutions, such as Hadoop, are changing this and allowing data of high volume, velocity, and variety to be much more easily analyzed. Hadoop is a massively parallel technology designed to be cost effective by running on commodity hardware.
Sentiment analysis isn't a panacea yet. It is still in its infancy, and there are limitations. "Despite significant advances in machine learning, it's extremely difficult (or not practically efficient) for computers to understand and process natural language, automate sentiment analysis, or determine ambiguous context," Wired stated just last month. "No matter how smart or stupid computers may be, it's just easier to create systems that encourage users to do their own interpretative work." But the technology is evolving across many industries, including high frequency trading, which relies on algorithms to execute market orders – and make profits – within miniscule fractions of a second.[19]

Tools used for Sentiment Analytics

1. Amazon EMR
2. SQL Server Integration Services
3. Apache Flume
4. DataSift
5. Hadoop
6. PowerPivot
7. PowerView
8. Sqoop
9. SQL Server Integration Services
10. and Polybase (SQL Server PDW 2012 only)
11. HDFS, HCatalog, Hive.
12. SAP Hana

4.4 Brand Monitoring / Price Comparison

Brand monitoring uses to get customer experience, what is best choice of customer with timing periods. For ex. Which colour is in demand for customer? Some companies uses other competitors data to get choice of customer what they want?
This analysis only works for "big Brands" with sufficient search volume. You can have your brand, products or services and compared with your competitor's in different regions and different time. The result will give you a glimpse on the "marketing share" in a particular region in particular time.
Crawling social media sites for extracting information is a fairly new concept - mainly due to the fact that most of the social media networking sites have cropped up in the last decade or so. But it's equally (if not more) important to grab this ever-expanding User-Generated-Content (UGC) as this is the data that companies are interested in the most - such as product/service reviews, feedback, complaints, brand monitoring, brand analysis, competitor analysis, overall sentiment towards the brand, and so on.
Scraping social networking sites such as Twitter, Linkedin, Google Plus, Instagram etc. is not an easy task for in-house data acquisition departments of most companies as these sites have complex structures and also restrict the amount and frequency of the data that they let out to crawlers. This kind of a task is best left to an expert, such as PromptCloud's Social Media Data Acquisition Service - which can take care of your end-to-end requirements and provide you with the desired data in a minimal turnaround time.
Most of the popular social networking sites such as Twitter and Facebook let crawlers extract data only through their own API (Application Programming Interface), so as to control the amount of information about their users and their activities.

Tools:
Google trends

Chapter 5
Models in Big data

5.1 Predictive Models

Predictive modeling is used to create a statistical model for future behavior. It's also used in email filtering systems to identify the probability that a given message is spam.
Predictive modeling is a process used in predictive analytics to create a statistical model of future behavior. Predictive analytics is the area of data mining concerned with forecasting probabilities and trends.
A predictive model is made up of a number of predictors, which are variable factors that are likely to influence future behavior or results. In marketing, for example, a customer's gender, age, and purchase history might predict the likelihood of a future sale. [20]

In predictive modeling, data is collected for the relevant predictors, a statistical model is formulated, predictions are made and the model is validated (or revised) as additional data becomes available. The model may employ a simple linear equation or a complex neural network, mapped out by sophisticated software.

Predictive modeling is used widely in information technology (IT). In spam filtering systems, for example, predictive modeling is sometimes used to identify the probability that a given message is spam. Other applications of predictive modeling include customer relationship management (CRM), capacity planning, change management, disaster recovery, security management, engineering, meteorology and city planning.

5.2 Prescriptive Models

Prescriptive analytics is the third and final phase of business analytics (BA) which includes descriptive, predictive and prescriptive analytics.

Referred to as the "final frontier of analytic capabilities," Prescriptive analytics automatically synthesizes big data, multiple disciplines of mathematical sciences and computational sciences, and business rules, to make predictions and then suggests decision options to take advantage of the predictions. The first stage of business analytics is descriptive analytics, which still accounts for the majority of all business analytics today. Descriptive analytics answers the questions what happened and why did it happen. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Most management reporting - such as sales, marketing, operations, and finance - uses this type of post-mortem analysis. [21]

Prescriptive Analytics extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes, and the interrelated effects of each decision
The next phase is predictive analytics. Predictive analytics answers the question what will happen. This is when historical performance data is combined with rules, algorithms, and occasionally external data to determine the probable future outcome of an event or the likelihood of a situation occurring. The final phase is prescriptive analytics, which goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the implications of each decision option.

Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen. Further, prescriptive analytics suggests decision options on how to take advantage of a future opportunity or mitigate a future risk and shows the implication of each decision option. Prescriptive analytics can continually take in new data to re-predict and re-prescribe, thus automatically improving prediction accuracy and prescribing better decision options. Prescriptive analytics ingests hybrid data, a combination of structured (numbers, categories) and unstructured data (videos, images,

sounds, texts), and business rules to predict what lies ahead and to prescribe how to take advantage of this predicted future without compromising other priorities.

All three phases of analytics can be performed through professional services or technology or a combination. In order to scale, prescriptive analytics technologies need to be adaptive to take into account the growing volume, velocity, and variety of data that most mission critical processes and their environments may produce.

One criticism of prescriptive analytics is that its distinction from predictive analytics is ill-defined and therefore ill-conceived.

5.3 Bayesian Model

Bayesian Model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a directed acyclic graph (DAG). For example, a Bayesian network could represent the probabilistic relationships between diseases and symptoms. Given symptoms, the network can be used to compute the probabilities of the presence of various diseases. [22]

Chapter 6
Benefits of Big data - Success Due to Big data

1. Aadhar Card

Everything about India's UID project or Aadhar as it is commonly known is ambitious. Giving a unique identity to 1.2 billion residents is a challenging task. No country has done a project of this scale – which is why this project is being watched keenly by everyone - not only in India, but the rest of the world too.

Another noteworthy part about the Aadhar project is the fact that the UID team wants to ensure that every record is indeed unique. To ensure that this happens, before every new ID issued, it is checked against the existing database. This process, which is called 'de-duplication' is done to ensure that no person gets more than one identity number. This is where it gets even more interesting. Till date, the Unique Identification Authority of India (UIDAI) has issued approximately 25 crore UIDs. Also, on a daily basis, the UIDAI issues over 1 million Aadhar cards. This means that each day, 1 million records have to be checked against the existing database of 25 crore IDs. This is only going to get larger every day. Additionally, as the UID database can be used for verifying identification, it receives more than 100 million authentications per day. Going forward, as more government-based systems access this platform to authenticate their customers or stakeholders, the number of authentication-based queries will also shoot up exponentially.

- 200 trillion biometric matches per day
- Peta Byte of raw data stored
- 100 million authentication requests per day
- Tera-byte scale data warehouse of 200 million records
- 50 million messages per day
- 100 million database transactions per day

Used Technologies: [**23**]

- Hadoop stack : HDFS, HBase, Hive, Pig, Zookeeper
- MySQL : sharded, partitioned, distributed
- SEDA : Mule, RabbitMQ
- Search : MongoDB, sharded Solr
- Compute Grid : Spring, GridGain
- Monitoring : Custom built, Nagios
- Analytics & Visualization
- Deployment footprint : Thousands of CPU cores
- Extensive Data archival, DR

2. Indian Election 2014

Hydrabad based Modak Analytics, a data analytics startup, built India's first Big Data – based electoral data repository of 81.4 crore voters for the just concluded elections. They processed around 81.4 crore electoral rolls. Undoubtedly the Indian election is the mother of all elections. Apart from the volume, we had 12 languages and this data changed very frequently.

3. Political Party Success – FB, Twitter Feed

Nowadays, every political persons are using Big data to get success, the best example. Mr. Narendra modi is first PM who used Big data.

"If you move out of the BJP website and visit a website for bikes followed by a search on jobs, the algorithm will make the inference that you are a young male from a particular

constituency, say Delhi, who is currently on a job hunt. What happens next is when you visit a job searching portal like Naukri.com, this system pops up a contextual ad for you like 'jobs in Delhi'. The BJP banner which is just below the results will tell you 'There are no Jobs in Delhi. India deserves better'." [24]

4. **The obama administration is investing $200 in big data research projects.**
5. **Police Crime detection**

London police uses big data for small crime detection. [25]

6. Traffic Control System

Ahmedabad Police uses Big data to solve traffic congestion problem to giving smart phone with installed app to every police who is on duty.

7. Importance in sports

Massive amount of data, along with hadoop and visualization tools helped the US woman's cycling team earn silver medal at London 2012 Olympic Games. [26]

8. NSE(Newyork Stock Exchange)

- New York Stock Exchange produces 1 TB data per day. [27]
- Call Center – Point of sale in Call center uses visualization and analytics.
- Sensors in health care and in auto car handled by big data hadoop.
- Farmers are getting climate change information with help of Big data.

9. Telco

Vodafone is reaching out to customers at the right time with the right offer, As Indian telecom space is hyper-competitive, telcos need to come up with innovative ways to attract customers and gain an edge over the competitors. One of the major challenges faced by telcos is choosing the right time to send the promotion offers to customers as they tend to forget the deals communicated to them by the time they walk into the outlet. In order to increase its promotions uptake and in turn improve its bottom-line, Vodafone India embarked on an innovative initiative, which aimed at reaching out to customers at the right time with the right offer.

Vodafone conducted a market research to understand recharge behavior, influence of retailer on choice of recharge, and the time a customer spends to get his/her mobile recharged at the Multi-Brand Outlets (MBOs). In line with outcome of the research, a best deal offer solution was developed to capture request from E-top up server on a real time basis, to send promotions within 30 seconds from the time of e-top-up trigger by the retailer. The solution was designed and interfaced with Etop-up solution, campaign management system, DND scrubbing system, push USSD gateway, and Short Message Service Center (SMSC). Interface was deployed to interact with Campaign management system to fetch the subscriber specific offer.

The initiative has enhanced the revenue margins for Vodafone -- with a spent of Rs 4 million on IT solutions, additional annual revenues to the tunes of Rs 20 million per annum is being realized due to this initiative.[28]

10. E-Commerce: Amazon /Netflix are earning good to do Click Stream Analysis for attract customers.

11. Weather data: a large size of satellite image is analyzes by Big data for weather information.

12. Travelling – Hotels: Travelling and hotels website is uses customers tracking data to get valuable customer and to give right offer to customer.

Chapter 7
Opportunity & Scope of Big data

By 2018, the United States alone could face a shortage of 140,000 to 190,000 people with deep analytical skills as well as 1.5 million managers and analysts with the know-how to use the analysis of big data to make effective decisions - McKinsey Global Institute.

46% of organizations cite inadequate staffing or skills for big data analytics - TDWI Research.

More than three-fourths of 169 executives surveyed say staffing and training issues are the greatest obstacles to making the most of big data - Ventana Research.

While the majority of executives (58%) believe finding the right technology is the biggest challenge their companies face in analyzing data, the majority (56%) of IT decision-makers charged with implementing Big Data programs believe finding the right staff is a bigger challenge than finding the right technology – Avanade [29].

83% of data scientists surveyed felt that new technology would increase the demand for data scientists, and 64% believe that it will outpace the supply of available talent – EMC [30].

Today there is a shortage of trained Big Data technology experts, in addition to a shortage of analytics experts. This labor supply constraint will act as an inhibitor of adoption and use of Big Data technologies, and it will also encourage vendors to deliver Big Data technologies as cloud-based solutions – IDC [31].

There will be a 24% increase in demand for professionals with management analysis skills over the next eight years. The need for this specialized talent is being fuelled by an increased use of business analytics by companies to better understand the explosion of data - U.S. Bureau of Labor Statistics [32].

Already Big data has challenge and issues as lacking of Data Science skills.

Figure 7.1 **:** Increasing Demand and Lack of Big Data Skills.
Source: http://www.edureka.in/blog/increasing-demand-for-hadoop-and-nosql-skills/

Chapter 8
Open Issues & Challenges - in Big data

8.1 Data Volume

Volume of data is a biggest challenge, as per size of data is increases companies are going to store data on different clouds. The data type that increases most rapidly is unstructured data. This data type is characterized by "human information" such as high-definition videos, movies, photos, scientific simulations, financial transactions, phone records, genomic datasets, seismic images, geospatial maps, e-mail, tweets, Facebook data, call-center conversations, mobile phone calls, website clicks, documents, sensordata, telemetry, medical records and images, climatology and weather records, log files, and text.
According to ComputerWorld, unstructured informationmay account for more than 70% to 80%of all data in organizations. These data, which mostly originate from social media, constitute 80% of the data worldwide and account for 90% of Big Data. Currently, 84% of IT managers process unstructured data, and this percentage is expected to drop by 44% in the near future. Most unstructured data are not modeled, are random, and are difficult to analyze. For many organizations, appropriate strategies must be developed to manage such data. Table 1 describes the rapid production of data in various organizations further.
According to Industrial Development Corporation (IDC) and EMC Corporation, the amount of data generated in 2020 will be 44 times greater [40 zettabytes (ZB)] than in 2009. This rate of increase is expected to persist at 50% to 60% annually. To store the increased amount of data, HDDs must have large storage capacities. Therefore, the following section investigates the development rate of HDDs.

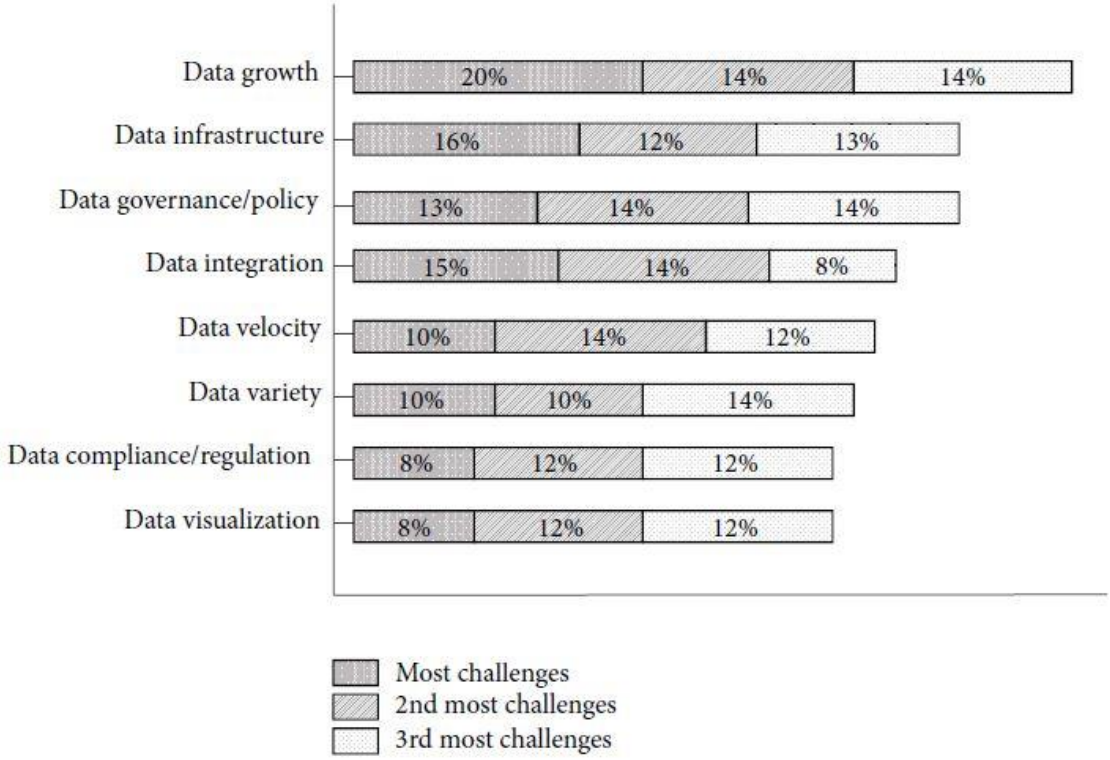


Figure 8.1: Volume - Big data Challange

8.2 Pre – Processing

Unstructured data is always challenging task for pre-processing. To Cleaning data, to remove missing value and noisy value, to do data quality, data integration, data migration, data extraction, data transformation, data loading etc are tough task.

8.3 Performance of CPU

To process very huge amount of unstructured data we requires very high configuration CPU and hardware's, cost is a challenge for companies where a large no. of CPU consumes a lot power energy as well as generates high heat. CPUs, memory, and disks are all subject to rapid innovation. Over time, fundamental shifts occur in the most efficient use of the available technologies.

8.4 Privacy & Security

Unfortunately, legitimate organizations aren't the only groups that are going big. Large sets of consolidated data are a tempting target for cyber attackers. Breaching an organization's big data repository can provide criminal groups with bigger payoffs and more recognition from a single attack. And when attackers set their sights on big data repositories, the effects can be devastating for the affected organizations. Terabytes of data in these repositories may include a company's crown jewels: customer data, employee data, and trade secrets. The recent data breach at Target is estimated to cost the company upwards of $1.1 billion, and the PlayStation breach cost Sony an estimated $171 million. A breach in a big data repository could be even more damaging at a financial institution or healthcare provider, where the value of the data is extremely high and government regulations come into play.
Securing big data comes with its own unique challenges beyond being a high-value target. It's not that big data security is fundamentally different from traditional data security. Big data security challenges arise because of incremental differences, not fundamental ones. The differences between big data environments and traditional data environments include: The data collected, aggregated, and analyzed for big data analysis The infrastructure used to store and house big data The technologies applied to analyze structured and unstructured big data.

Securing Big Data

So what can be done to help bring the security of traditional database management to big data? Several organizations describe and define different security controls. The SANS Institute provides a list of 20 security controls. The list contains several controls that I would recommend to address the security challenges presented by big data.

Application Software Security.Use secure versions of open-source software. As described above, big data technologies weren't originally designed with security in mind. Using open-source technologies like Apache Accumulo or the .20.20x version of Hadoop or above can help address this challenge. In addition, proprietary technologies like Cloudera Sentry or DataStax Enterprise offer enhanced security at the application layer. Specifically, Sentry and Accumulo also support role-based access control to enhance security for NoSQL databases.

Maintenance, Monitoring, and Analysis of Audit Logs. Implement audit logging technologies to understand and monitor big data clusters. Technologies like Apache Oozie can help implement this feature. Keep in mind that security engineers in the organization need to be tasked with examining and monitoring these files. It's important to ensure that auditing, maintaining, and analyzing logs are done consistently across the enterprise.

Secure Configurations for Hardware and Software. Build servers based on secure images for all systems in your organization's big data architecture. Ensure patching is up to date on these machines and that administrative privileges are limited to a small number of users. Use

automation frameworks, like Puppet, to automate system configuration and ensure that all big data servers in the enterprise are uniform and secure.

Account Monitoring and Control. Manage accounts for big data users. Require strong passwords, deactivate inactive accounts, and impose a maximum permitted number of failed log-in attempts to help stop attacks from getting access to a cluster. It's important to note that the enemy isn't always outside of the organization. Monitoring account access can help reduce the probability of a successful compromise from the inside.

Organizations that are serious about big data security should consider these first steps. Cyber criminals are never going to stop being on the offensive, and with such a big target to protect, it is prudent for any enterprise utilizing big data technologies to be as proactive as possible in securing its data.

8.5 Lack of Skill / Talent

Big data is now buzzword and rapid growth of data science already generated big demand of data scientist that diverts people to pursue data science degree.
Recently, UC Barkley started to offer Online Data science degree with $ 60,000 [33] , In the past few years, as data science has become the "sexiest" Job of the century, other top universities, like North-western and New York University, have moved into this area.
MIT also offers online Big data degree ("Tackling the Challenges of Big Data") with $ 545 [34].
All above online degree cost a lot to people, so people not having really good financial condition cannot afford the high amount of cost to pursue big data course, and at other side Big data and Data science is sexiest job of 21st century [35] so everybody wants to become big data expert and data scientist, Gartner Says Big Data Creates Big Jobs: 4.4 Million IT Jobs Globally to Support Big data by 2015 [36].
So this stage must confuse people a lot that which is best path to do? Solution is Open Couseware (MOOC), researchers and students believe that MOOC (Massive Open Online Course) is golden boon for Research literacy, Open Courseware that can help people to become big data expert, Data scientist. As above high cost Online degrees from Berkeley, MIT requires computer science knowledge, MOOC also requires some sense of Computer Science knowledge.
Who offers Open Courseware programs?

- Edx.org
- Coursera.org
- Udacity.com

  etc.

Above 3 are most popular portals to accomplish Open Courseware program from well knowledgeable professors at top reputed University of world such as Harvard University, Stanford University, MIT University, IIT etc.

8.6 Challenges in Big data projects

Based on rapid growth of Unstructured data, big data increasing its challenges.

Sources of Big data

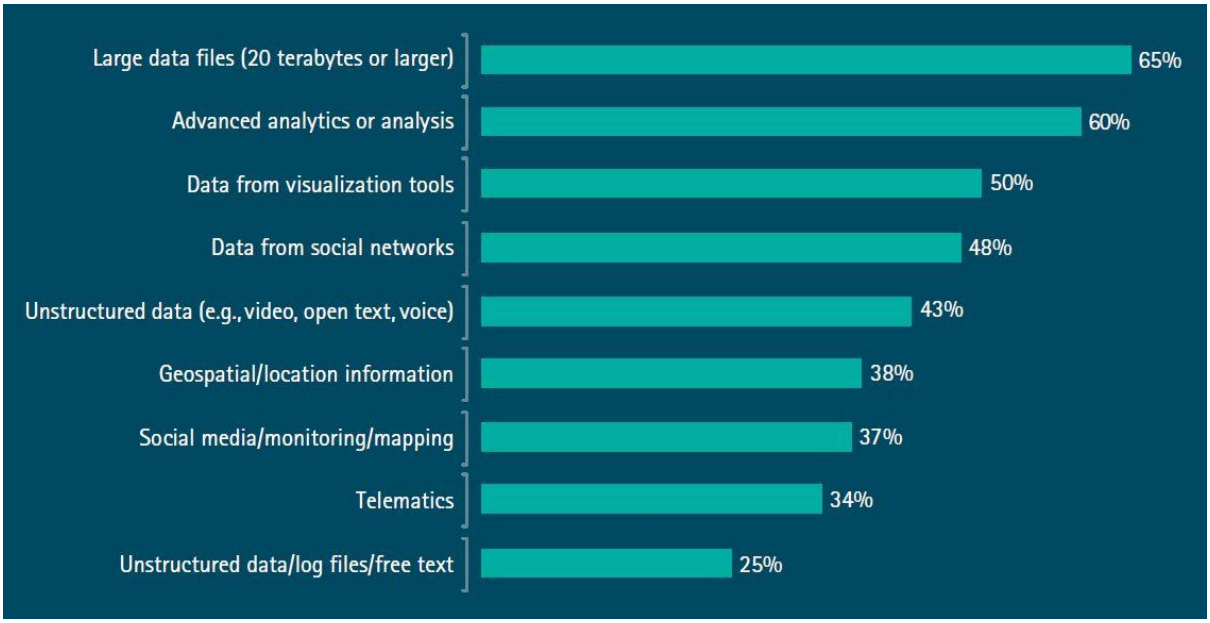


Figure 8.2 : Sources of big data

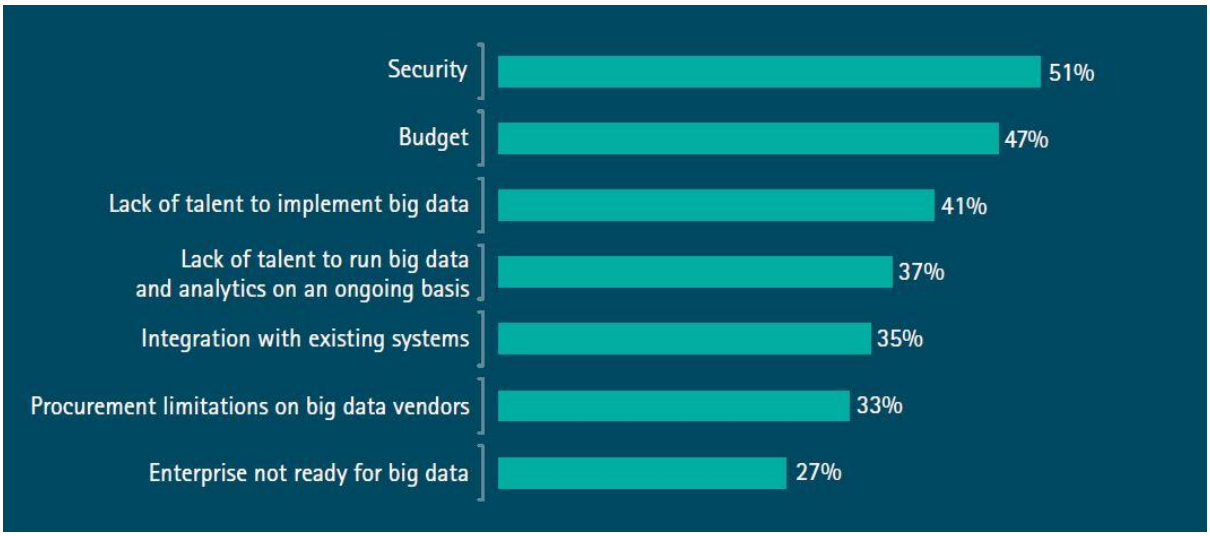


Figure 8.3 : Main challenges of Big data for Companies

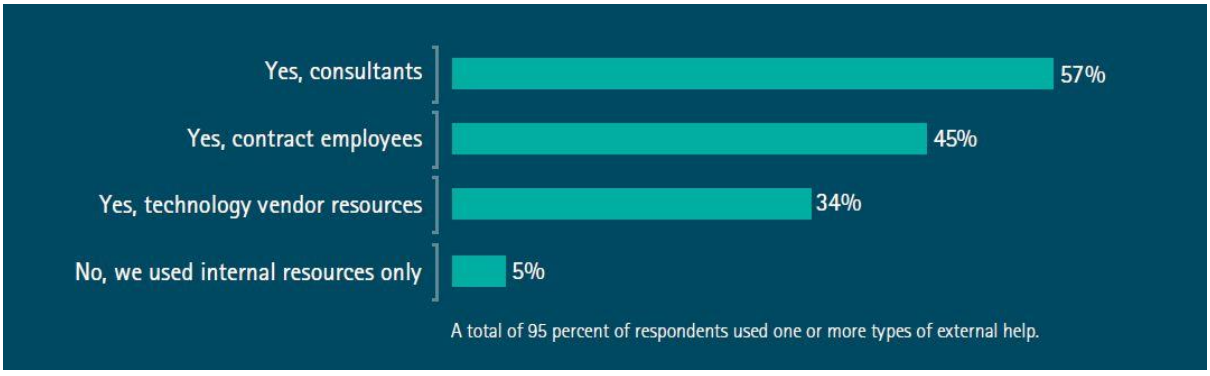


Figure 8.4 : External help for Big data installation

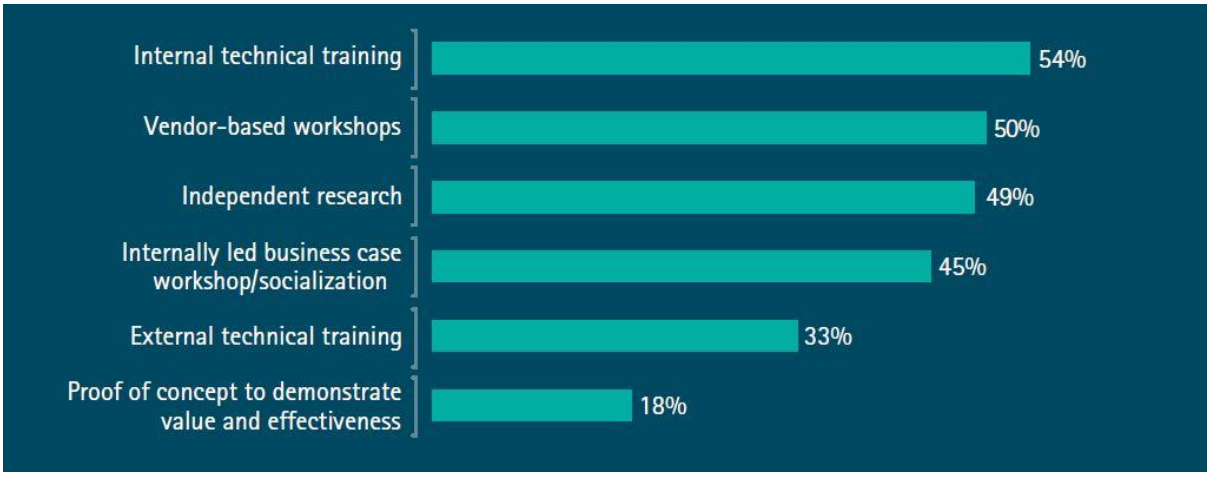


Figure 8.5: Addressing big data challenges

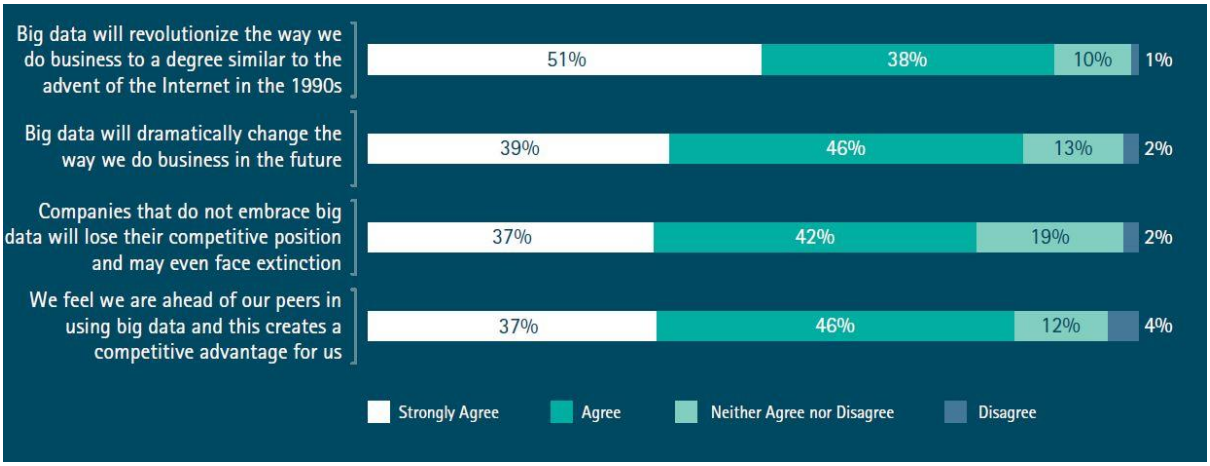


Figure 8.6: Big data's competitive significance

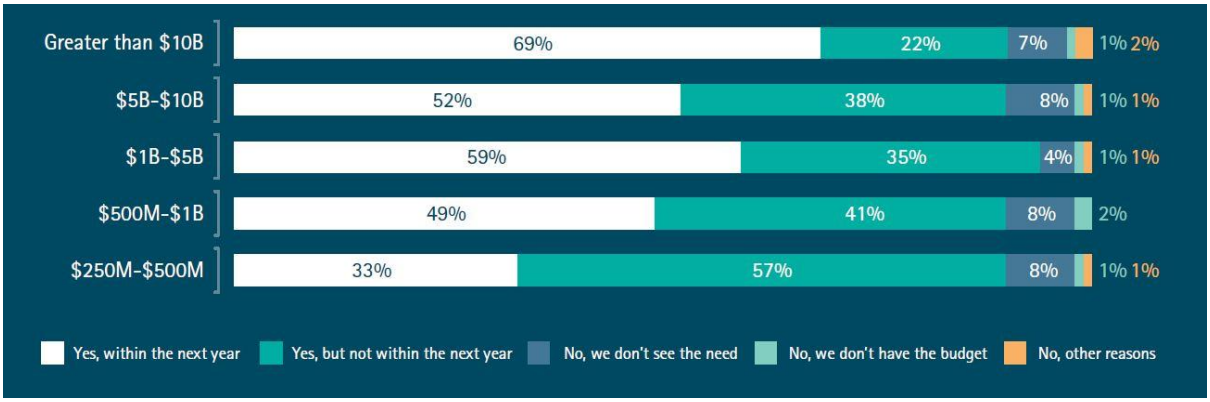


Figure 8.7: Big data Investment in near term

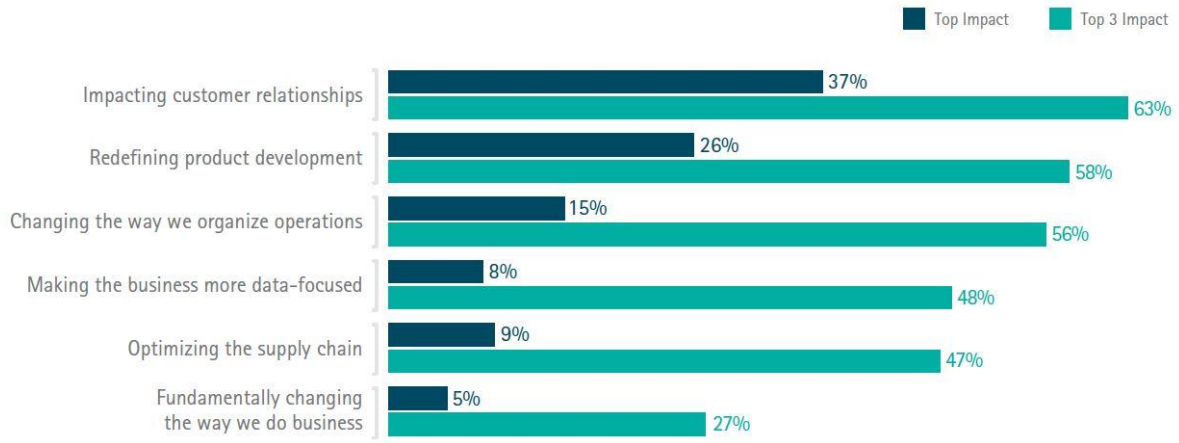


Figure 8.8: Big data impact in next five years

With so many organizations simultaneously competing for big data skills, sourcing talent is undeniably difficult.

- More than half of respondents (57 percent) leveraged the help of consultants, 45 percent used contract employees and 34 percent used technology vendor resources (see Figure 4).
- Organizations that relied on consultants, contractors and other external resources found their big data installations to be easier than those using only internal resources. The big data skills shortage is likely to persist in the near term, making this one problem companies cannot overcome through hiring alone. To address the talent shortage crisis and other challenges, companies resort to a spectrum of strategies (see Figure 5).
- Nearly all (91 percent) companies expect to increase their data science expertise, the majority within the next year.
- Training, workshops and research are used to address the talent challenge by developing skills internally. Successful big data practitioners are leveraging big data and big data technologies to drive business outcomes. An outcome focus requires an ability to mobilize data from across the enterprise; to interrogate that data deeply to understand its value, and determine what data is important and what data is not; and requires a discipline to govern it so that it maintains its currency in the enterprise.
- As more data is available, it demands to be quantified, quickly. New methods and approaches for data discovery mean analytic driven insights are generated in weeks or months. Agile approaches are employed to drive rapid, demonstrable progress.

Working with big data necessarily places companies in a sphere that is potentially rich with inadvertent discovery and innovation. Understanding business use cases and data usage patterns (the people and things that consume data) provides crucial evidence into the appropriate solutions, technologies and approaches that will be used to deliver results. Multiple solutions exist for any given big data challenge, so it is vital to remain open to the possibilities, and become a learning enterprise by testing extensively, learning what works best, then refining and moving forward. Big data pioneers have honed their capacity to test everything and learn quickly; other companies are emulating these practices.

Chapter 9
Bibliography

[1] IBM: Wrangling big data: Fundamentals of data lifecycle management, How to maintain data integrity across production and archived data.

[2] Robin Bloor, Ph.D. & Rebecca Jozwiak (The Bloor Group) THE BIG DATA INFORMATION ARCHITECTURE An Analysis of the Consequences of the Big Data Trend, p.p 18 – 29

[3] Getting Big Value From Big Data... Fast An ENTERPRISE MANAGEMENT ASSOCIATESff (EMA™) White Paper Prepared for Tableau Software May 2011 p.p 1

[4] [Online]. Available. Facebook, Facebook Statistics, 2014, http://www.statisticbrain .com/facebook-statistics/.

[5] [Online]. Available. YouTube, "YouTube statistics," 2014, http://www.youtube.com/ yt/press/statistics.html.

[6] [Online]. Available. Twitter, "Twitter statistics," 2014, http://www.statisticbrain .com/twitter-statistics/.

[7] [Online]. Available. Foursquare, "Foursquare statistics," 2014, https://foursquare .com/about.

[8] [Online]. Available. Jeff Bullas, "SocialMedia Facts and Statistics You Should Know in 2014," 2014, http://www.jeffbullas.com/2014/01/17/20-socialmedia-facts-and-statistics-you-should-know-in-2014/.

[9] [Online]. Available. Marcia, "Data on Big Data," 2012, http://marciaconner.com/ blog/data-on-big-data/.

[10] [Online]. Available. Taming Big Data [A Big Data Infographic] by Wikibon.org http://wikibon.org/blog/taming-big-data/

[11] [Online]. Available. What is apache hadoop http://hadoop.apache.org/

[12] [Online]. Available. Apache Hadoop http://en.wikipedia.org/wiki/Apache˙Hadoop

[13] [Online]. Available. MapReduce http://en.wikipedia.org/wiki/MapReduce

[14] Big Data Gold Accreditation: Stan Dorcey Adam Bracey Informatica Conference 2014

[15] [Online]. Available. NOSQL : http://en.wikipedia.org/wiki/NoSQL

[16] [Online]. Available. Data stream management system http://en.wikipedia.org/wiki/Data˙stream˙management˙system

[17] [Online]. Available. Real-Time Stream Processing as Game Changer in a Big Data World with Hadoop and Data Warehouse http://www.infoq.com/articles/stream-processing-hadoop

[18] [Online]. Available. Sentiment Analysis with Big Data SAP:
http://www.news-sap.com/sentiment-analysis- with-big-data/

[19] [Online]. Available. Sentiment Analysis by AWS
http://docs.aws.amazon.com/gettingstarted/latest/emr/getting-started-emr-sentiment-tutorial.html

[20] [Online]. Available. Predictive Modeling
http://searchdatamanagement.techtarget.com/definition/predictive-modeling

[21] [Online]. Available. Prescriptive Analytics
http://en.wikipedia.org/wiki/Prescriptive˙analytics

[22] [Online]. Available. Bayesian network
http://en.wikipedia.org/wiki/Bayesian˙network

[23] [Online]. Available. Aadhaar - world's largest biometric identity platform (200 trillion biometric matches per day, 2 PB of data)
https://fifthelephant.talkfunnel.com/2012/417-aadhaar-worlds-largest-biometric-identity-platform-200-trillion-biometric-matches-per-day-2-pb-of-data

[24] [Online]. Available. The Big Secret behind Narendra Modi's win
http://www.venturesity.com/blog/the-big-secret-behind-narendra-modis-win

[25] [Online]. Available. London police using Big Data to tackle small crime
http://www.cloudcomputing-news.net/news/2014/oct/31/london-police-using-big-data-tackle-small-crime/

[26] [Online]. Available. Can Big Data Trump Doping In Sports?
http://www.informationweek.com/big-data/big-data-analytics/can-big-data-trump-doping-in-sports/d/d-id/1110593?

[27] Hadoop: The Definitive Guide By tom white

[28] Telco - http://www.informationweek.in/informationweek/case-study/298153/vodafone-reaching-customers-offer

[29] [Online]. Available. Ease Big Data Hiring Pain with Cascading — CIO
www.cio.com/article/.../ease-big-data-hiring-pain-with-cascading.html

[30] [Online]. Available. Intent to Plan: M.S. in Analytics and M.S. in Data Science Dakota State University and
South Dakota State University.
www.sdbor.edu/theboard/agenda/2013/December/CommA/III˙A.pdf

[31] IDC Releases First Worldwide Big Data Technology and Services Market Forecast, Shows Big Data as the Next Essential Capability and a Foundation for the Intelligent Economy.
[Online]. Available. http://www.businesswire.com/news/home/20120307005036/en/IDC-Releases-Worldwide-Big-Data-Technology-Services

[32] Analytics Goes to the Head of the Class: Northwestern University News
http://www.northwestern.edu/newscenter/stories/2011/12/ibm-analytics-masters.html

[33][Online]. Available. http://datascience.berkeley.edu/

[34] [Online]. Available.

https://mitprofessionalx.edx.org/courses/MITProfessionalX/6.BDX/2T2014/about#overviw

[35] [Online]. Available. https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century/

[36] [Online]. Available. http://www.gartner.com/newsroom/id/2207915