Big Data and Data Science: Case Studies

Priyanka Srivatsa[1]

[1]Department of Computer Science & Engineering, M.S.Ramaiah Institute of Technology, Bangalore- 560054.

***Abstract*- Big data is a collection of large and complex data sets difficult to process using on-hand database management tools or traditional data processing applications. The three V's of Big Data (volume, variety, velocity) constitute a more comprehensive definition busting the myth that big data is only about data volume. Data Volume is the primary attribute of big data. It can be quantified by counting records, transactions, tables or files. It can also be quantified in terms of time & in terms of terabytes or petabytes. Data Variety, the next significant attribute of big data, is quantified in terms of sources like logs, clickstream or social media. Data velocity, another important attribute of big data, describes the frequency of data delivery & data generation. Analysis of big data is very complex & time consuming. An important tool that helps understand big data & its analysis is Data Science. Data Science is the study of the generalizable extraction of knowledge from the data sets. The study of data science includes studying data processing architectures, data components & processes, data stores & data kind and the challenges of big data.**

Keywords:-big data, volume, variety, velocity, data science, data processing pipelines, data processing architectures, data components, data processes, data stores, data kind

I. INTRODUCTION

Data processing pipeline typically has 5 phases:

1. Data Acquisition & Recording - Data is recorded from various sources.
2. Information Cleaning & Extraction - The required information is extracted from the underlying processes & expressed in a structured form suitable for analysis.
3. Data Integration, Aggregation & Representation - Differences in data structure & semantics need to be identified & understood and an intelligent database design is developed making data computer understandable.
4. Query Processing, Data Modelling & Analysis -Big data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Interconnected Big data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to cross-check conflicting cases, to validate trustworthy relationships, to disclose inherent clusters and to uncover hidden relationships & models. Data Mining also helps improve the quality & trustworthiness of data, understanding its semantics & thus provides intelligent querying functions.
5. Interpretation: Ultimately, the decision-maker, provided with the result of analysis, has to interpret the result which involves examining all the assumptions made & retracing the analysis to check the possible sources of error: bugs in the computer system, assumptions of the models & results based on erroneous data.

Data Components provides access to data hosted within the boundaries of the system.

Data Processes are those processes that help in collection & manipulation of meaningful data.

Data Store is a data repository of a set of integrated objects. These objects are modeled using classes defined in database schemas.

Data kind refers to the variety of data available for analysis. It includes structured data, unstructured data & semi-structured data. Structured Data exists when the information is clearly broken down into fields that have an explicit meaning & are highly categorical, ordinal or numeric. Unstructured Data exists in the form of natural language text, images, audio & video. It requires pre-processing to identify & extract relevant features. Semi-Structured Data is used to describe the structured data that does not conform to the formal structure of data models associated with a relational database or other forms of data tables.

Challenges in Big Data Analysis are:

1. Heterogeneity & Incompleteness: Machine analysis algorithms expect homogeneous data & cannot understand nuance. Even after data cleaning and error correction, some incompleteness & errors in data are likely to remain.
2. Scale: Managing large & rapidly increasing volume of data has been a big challenge for the past few decades.
3. Timeliness: The larger the data set, the longer it will take to analyze.
4. Privacy: Managing privacy is effectively both a technical & a sociological problem that must be addressed jointly from both perspectives to realize the promise of big data.
5. Human Collaboration: The data system needs to be designed such that it accepts the distributed expert input & supports their collaboration.

This paper discusses the data processing pipelines, data stores, data components, data processes, data kind & the challenges of big data analytics in a real-world scenario.

II. CASE STUDY - I

Problem Statement:

Nokia, a top telecom firm, has a goal; to bring the world to the third phase of mobility: leveraging digital data to make it

easier to navigate the physical world. To achieve this goal, Nokia needed to find a technology solution that would support the collection, storage and analysis of virtually unlimited data types and volumes. Effective collection and use of data has become central to Nokia's ability to understand and improve users' experience with their phones. The company leverages data processing and complex analyses in order to build maps with predictive traffic and layered elevation models, to source information about points of interest around the world & understand the quality of phones. Cloudera helped Nokia in its endeavor to achieve this goal by deciding to employ APACHE HADOOP to manage & process huge volumes of data.

A. Data Processing Architecture

Data required for analysis is acquired from various resources like phones in use, services, log files, market research, discussion in forums, feedback etc. All this data is sent into a DATA COLLECTOR which collects & stores these various kinds of data required for analysis. After initial data collection, a cleaning process is conducted with sampling & conversion of data.

Then the data is aggregated & sent into a DATA PROCESSOR. This complete process is supervised by a DATA SUPERVISOR that appropriately pre & post processes the live data. The aggregated data is sent into a COMPUTE CLOUD component that consists of 3 parts namely the Data Broker, the Data Analyzer & the Data Manager.

1).Data Broker collects & repackages information available in the public domain in a format readable & useful to the company.

2).Data Analyzers are tools that specialize in predictive modeling & text mining thus analyzing the information available.

3).Data Manager is a tool that manages the processing of huge volumes of data by realizing the entities of applications & efficiently creating graphs & information snapshots that deliver the analysis into a presentable format.

There is a DATA REST unit that consists of:

a).QUERY unit used to query from the database

b).REPORTER unit that reports the results of the related queries

c).CACHE unit that stores all the temporary information retrieved from the database

d).VISUALIZER unit that helps process the digital data & interpret results

e).AUDIT unit that keeps an account of the amount of data that is being processed & the effective time required to process this data

f).MONITOR unit which monitors the entire functioning of the DATA REST unit

The COMPUTE CLOUD component is supported by the DATA REST unit.

Finally the processed data is fed into the DATA SAAS wherein the various dimensions & prospects of the data are discussed & interpreted.

B. Components

The technology ecosystem consists of:

1).Teradata Enterprise Data Warehouse: It stores & manages data.

2).Oracle & My SQL Data Marts: These are simpler forms of data warehouses.

3).HBase: It is an extensible record store with a basic scalability model of splitting rows & columns into multiple nodes.

4).Scribe: It is used to log data directly into the HBase.

5).Sqoop: It is a command-line interface application for transferring data between relational databases and Hadoop (HBase).

C. Process

1).Nokia has over 100 terabytes (TB) of structured data on Teradata and petabytes (PB) of multi-structured data on the Hadoop Distributed File System (HDFS).

2).The centralized Hadoop cluster which lies at the heart of Nokia's infrastructure contains 0.5 PB of data.

3).Nokia's data warehouses and marts continuously stream multi-structured data into a multi-tenant Hadoop environment, allowing the company's 60,000+ employees to access the data.

4).Nokia runs hundreds of thousands of Scribe processes each day to efficiently move data from, for example, servers in Singapore to a Hadoop cluster in the UK data center.

5).The company uses Sqoop to move data from HDFS to Oracle and/or Teradata.

6).And Nokia serves data out of Hadoop through HBase.

D. Data Stores

1).Teradata Enterprise Data Warehouse: This data warehouse uses a "shared nothing" architecture which means that each server node has its own memory and processing power. Adding more servers and nodes increases the amount of data that can be stored. The database software sits on top of the servers and spreads the workload among them.

2).Oracle & My SQL Data Marts: These are focused on a single subject (or functional area), such as Sales, Finance or Marketing. Data marts are often built and controlled by a single department within an organization. Given their single-subject focus, data marts usually draw data from only a few sources. The sources could be internal operational systems, a central data warehouse or external data.

3).HBase: HBase is an Apache project written in Java. It is patterned directly after Big Table:

• HBase uses the Hadoop distributed file system which updates memory and periodically writes them out to files on the disk.

• The updates go to the end of a data file, to avoid seeks. The files are periodically compacted. Updates also go to the end of a write ahead log, to perform recovery if a server crashes.

• Row operations are atomic, with row-level locking and transactions. There is optional support for transactions with wider scope. These use optimistic concurrency control, aborting the process; if there is a conflict with other updates.

• Partitioning and distribution are transparent; there is no client-side hashing or fixed key space. There is multiple master support, to avoid a single point of failure.

MapReduce support allows operations to be distributed efficiently.

• HBase's B-trees allow fast range queries and sorting.

• There is a Java API, a Thrift API and REST API; JDBC/ODBC support has recently been added.

E.Data Kind

1).Unstructured Raw Data: services, log files, images, customer feedback, discussion in various forums etc.
2).Structured Data: graphs, information snapshots etc.

F. Business Challenges that enabled this project:

1). Numerous groups inside Nokia were building silos to accommodate individual needs. The company realized that for effective understanding of the customer's needs, they needed to integrate all these individual silos into a single comprehensive data set.
2). Nokia wanted to understand at a holistic level how people interact with different applications around the world, which required them to implement an infrastructure that could support daily, terabyte-scale streams of unstructured data from phones in use, services, log files and other sources.
3). Leveraging this data also required complex processing and computation to be consumable and useful for a variety of uses, like gleaning market insights or understanding collective behaviors of groups; some aggregations of that data also need to be easily migrated to more structured environments in order to leverage specific analytic tools.
4). Capturing petabyte-scale data using a relational database was cost prohibitive and would limit the types of data that could be ingested.
5). Unstructured data had to be reformatted to fit into a relational schema before it could be loaded into the system. This required an extra data processing step that slowed ingestion, created latency and eliminated elements of the data that could become important down the road.

G. Comparative Study of the Use Case & the BIG DATA PIPELINE.

The Data Processing Pipeline we studied has 5 significant phases which Cloudera has incorporated in the DATA PROCESSING PIPELINE they designed for NOKIA.
The association is as follows:
a). Data Acquisition & Recording is a phase where data is acquired from the various sources. The Teradata Enterprise Warehouse acquires this continuous data for processing.
b). Information Extraction & Cleaning is a phase where the data is extracted according to the requirement & made analysis ready. Teradata Enterprise Warehouse extracts the data as per the requirement & makes it analysis ready.
c). Data Aggregation, Integration & Representation is a phase where relevant data for analysis is grouped considering the heterogeneity of the data acquired. The Oracle & My SQL data marts separate the relevant data effectively as data marts concentrate on concrete, single subjects specifically on one functional area.
d). Query Processing, Data Modelling & Analysis is a phase where general statistic patterns are drawn from hidden patterns. HBase effectively derives the statistic patterns from hidden patterns.
e).Interpretation is a phase wherein all the assumptions made need to be examined & the possible errors have to be removed. Data SaaS available in the Hadoop Framework are utilized to interpret the processed data results efficiently.

III. CASE STUDY - II

Problem Statement

A retail supplier and buyer of medical equipment with a growing customer base, product lines, partners and vendors needed a 360-degree view of its core business entities, transactional information and integrated data for business analytics.
The inability to get this critical information was prolonging customer and product management time, affecting overall time to market products.

A. Data Processing Architecture

The data processing pipeline for this retail supplier consists of 5 Phases:

a) Data Acquisition & Recording.

The retail supplier has a growing customer base, product lines, partners & vendors. This means that these pillars are the sources of critical information to understand the core business entities, the transactional information & the core of business analytics for this company. This information needs to be acquired & recorded for further data processing.

b) Information Extraction & Cleaning.

There is a huge volume of critical data being flooded for analysis. But this data is heterogeneous in nature & needs to be collected & converted into readable format. Information Extraction Process pulls out the required information from the underlying sources & expresses it in a structured form suitable for analysis.

c) Data Integration, Aggregation & Representation.

Given the heterogeneity of the flood of data, it is not enough to merely record & throw the data into a repository. Data analysis becomes a critical phase as it requires differences in data structure & semantics to be expressed in forms that are computer understandable & then robotically resolvable. Domain Scientists created effective database designs, either through devising tools to assist the design process or by developing techniques so that databases can be used effectively in the absence of intelligent data design.

d) Query Processing, Data Modelling & Analysis.

Big Data is often more noisy, dynamic, heterogeneous, inter-related & untrustworthy. General statistics obtained

from frequent patterns & correlation analysis usually overpower individual fluctuations & often disclose more reliable hidden patterns & knowledge. Interconnected Big Data forms large heterogeneous information networks, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent clusters and to uncover hidden relationships and models. Mining requires integrated, cleaned, trustworthy and efficiently accessible data, declarative query and mining interfaces, scalable mining algorithms and big-data computing environments. Data mining itself is being used to help improve the quality and trustworthiness of the data, understand its semantics and provide intelligent querying functions.

e) Interpretation.

Ultimately the results of analysis need to be interpreted by a decision maker. The process basically involves examining all the assumptions made & retracing the analysis. The errors have to be debugged & the assumptions at various levels need to be critically examined. Supplementary information that explains the derivation of each result & the inputs that are involved in this process need to be mentioned & explained wherever necessary.

B. Components

Oracle deployed its MDM suite that consists of the following components:

a) Oracle Metadata Manager: It acquires & records the continuous inflow of data into the database.
b) Data Relationship Manager: It consolidates, rationalized, governs & shares the master reference data.
c) Data Warehouse Manager: It divides the acquired data into specific functional areas & stores them in data marts.
d) BI Publisher: It queries, monitors & reports on the master data.
e) Data Steward Component: It facilitates the UI component & also helps to set up the workbench.

C. Process

1) Profile the master data. Understand all possible sources and the current state of data quality in each source.
2) Consolidate the master data into a central repository and link it to all participating applications.
3) Govern the master data. Clean it up, de- duplicate it, and enrich it with information from 3rd party systems. Manage it according to business rules.
4) Share it. Synchronize the central master data with enterprise business processes and the connected applications. Insure that data stays in sync across the IT landscape.
5) Leverage the fact that a single version of the truth exists for all master data objects by supporting business intelligence systems and reporting.

D. Data Stores

1) Siebel: It is used exclusively to store CRM data.
2) EBS: It provides persistent block-level storage volumes for use with Amazon EC2.
3) SAP: It serves as a storage location for consolidated & cleansed transaction data on an individual level.
4) JDE: It is used to provide periodic updates of the operational data changes required.
5) PSFT: It is used as a data store to manage entire business process relationships.

E. Data Kind

1) Transaction data are business transactions that are captured during business operations and processes, such as a purchase records, inquiries, and payments.
2) Metadata, defined as "data about the data", is the description of the data.
3) Master data refers to the enterprise-level data entities that are of strategic value to an organization. They are typically non-volatile and non-transactional in nature.
4) Reference data are internally managed or externally sourced facts to support an organization's ability to effectively process transactions, manage master data, and provide decision support capabilities. Geo data and market data are among the most commonly used reference data.
5) Unstructured data make up over 70% of an organization's data and information assets. They include documents, digital images, geo-spatial data, and multi-media files.
6) Analytical data are derivations of the business operation and transaction data used to satisfy reporting and analytical needs. They reside in data warehouses, data marts, and other decision support applications.
7) Big data refer to large datasets that are challenging to store, search, share, visualize, and analyze.
The growth of such data is mainly a result of the increasing channels of data in today's world.
Examples include, but are not limited to, user-generated content through social media, web and software logs, cameras, information-sensing mobile devices, aerial sensory technologies, genomics and medical records.

F. Business Challenges that enabled this project:

1) SUPPLY CHAIN MANAGEMENT was crucial to this retail supplier as he was facing loses because of unstructured data management.
2) The time he took to market his products was really huge which led to other companies marketing similar stuff.
3) His revenue decreased as his sales fell significantly due to his inability to manage the data.
4) SHIPPING & INVOICING ERRORS were huge that led to economic & customer losses.
5) Distribution slowed down owing to inadequate management of required data.
6) Errors in acquiring orders resulted in dissatisfaction of the customers.

G. Comparative Study of the Use Case & the BIG DATA PIPELINE.

a) Data Acquisition & Recording is a phase where data is acquired from the various sources. Oracle Metadata Manager Tool acquires the required data.
b) Information Extraction & Cleaning is a phase where the data is extracted according to the requirement & made analysis ready. Data Relationship Manager extracts the data as per requirement & makes it analysis ready.
c) Data Aggregation, Integration & Representation is a phase where relevant data for analysis is grouped considering the heterogeneity of the data acquired.
Data Warehouse Manager separates the relevant data effectively as data marts concentrate on concrete, single subjects specifically on one functional area.
d) Query Processing, Data Modelling & Analysis is a phase where general statistic patterns are drawn from hidden patterns. BI Publisher effectively derives the statistic patterns from hidden patterns.
e) Interpretation is a phase wherein all the assumptions made need to be examined & the possible errors have to be removed. Data Steward Component is utilized to interpret the processed data results to the workbench efficiently.

REFERENCES

[1] Big Data Analytics by Philip Russom
[2] Challenges and Opportunities with Big Data
[3] Scalable SQL & NoSQL Data Stores by Rick Catell
[4] Field Guide to Data Science by Booz, Allen & Hamilton
[5] An Architects' Guide to Big Data - Oracle white paper
[6] Cloudera - Nokia Case Study
[7] Oracle & Big Data White Paper
[8] Master Data Management by Oracle