AISHA HASSAN SHAH **45920842**

BIG DATA TECHNOLOGIES ITEC 874

ASSIGNMENT #1 DATA LAKE ARCHITECTURE

PART 1. DATA LAKE COMPONENTS

1. DATA INGESTION COMPONENT

A data lake is a humongous platform, which assimilates large volume and variety of structured, semi structure and unstructured data. This centralized repository has functionality of extracting data from any available data source and can be loaded inside it with different approaches such as Batch ingestion, Micro-batch ingestion and real-time ingestion.

A. (I) STRUCTURED DATA

Structured data strictly resides in relational databases (RDBMS) and follows the structures and constraints specific in a Schema. RDBMS is an application/system that allows data to be created, updated and maintained a relational database. SQL is the most usable language in relational database to access the database. SQL syntax may differ slightly in format while using in variety of RDBMS. **MySQL, PostgreSQL, Oracle DB, SQL Server, SQLite and many more are the example of Relational database management systems.**

A. (II) UN STRUCTURED DATA

Unstructured data is a data which is not organized in a predefined data models or any particular relational Schema and with very limited indication of the type of data. More than 70% of the data that is generated on daily basis are unstructured. It may be textual or non-textual, human or machine generated. It can also be stored and managed within a non-relational database like NoSQL. NoSQL databases are used to manage Semi Structured and Unstructured data.

For instance, Text files, Emails, Social Media, Website, Mobile data, Communications, Media, Business application and many more are human-generated unstructured data.

And machine-generated unstructured data includes Satellite imagery, scientific data, Digital surveillance, Sensor data and many more.

A. (III) SEMISTRUCTURED DATA

Semi structured data is a data which does not obey the formal structure of data models linked with relational databases or entity relationship schema. Each data in a semi structured database represents its own schema. It contains tags and elements (Metadata) to separate semantic data and enforces hierarchies on tables within the data. This structure may have missing attributes or same attribute repeating itself. Therefore it is also called self-describing structure. Semi-structured data exists in the middle of structured and unstructured data. It shows certain aspects that are structured, and others that are not.

Semi structured data can be stored in CSV, JSON and XML and other markup languages, binary executables ,zipped files, integration of data from different sources and many more.

A. (IV) DATA INGESTION

Data ingestion is a process in which data is imported/extracted from one or more sources to a destination where it can be stayed and stored for further procedures. The source data is available in different formats such as structured, unstructured and semi structured, since the source data comes from different platforms, it requires to be transformed, enriched and cleansed to be settled on destination(as an optimize data structures)for further steps in Data Lake. The data ingestion process takes place in different ways such as Batch ingestion, Micro-batch ingestion, Real time ingestion. When data is ingest in **batches**, it is imported (large data) at well-defined schedule time frame (sequential order) or any specific condition ordered to be triggered to execute the processing. **Real time** ingestion is the process, in which data is processed without time limitations (continuous computation), it is very time sensitive and each new entry of data is operated when it is arrived. Another is **Micro-batch** ingestion, which serves with small batches and processes in small intervals (even once every few minutes) but in a form of a batch at a time.

Note: Stream processing is the combine functionality of real time ingestion and Micro-batch ingestion.

1. Lambda architecture	• Lambda architecture is a technique of data processing that is capable of handling with both batch processing as well as real time processing. • Lambda architecture comprises with three different layers Batch layer, Speed layer (stream layer) and serving layer. These layers work by combining the outcomes of historical storage in the form of batches and real time (streaming) with the help of speed layer and providing solutions to queries on all data models. • It is used when the batch processing, micro-batch + real time (Stream processing) occurring and keep adding new entries of data to the main storage and ensuring that the existing data remains intact. Twitter, Netflix, and yahoo are the organization utilizing Lambda architecture to meet the service standards.
2.Kappa architecture	• Kappa architecture is the architecture which serves in real-time processing of different processing enterprise data models. • This architecture is functioned having two layers speed layer (streaming layer) and service layer. All queries can be answered by applying Kappa functions. • Kappa architecture can be used for those data processing enterprise models where order of the data generation and queries are not predetermined and stream processing platforms can correspond with database at any time.

B. There are many preferable selection of Big Data Technologies and Tools to cater with data ingestion. **Wavefront, Hortonworks Data Platform, Amazon Kinesis, Apache Kafka** and many more .Few of them are as follows.

1.Wavefront	• Wavefront is a query engine and a hosted platform serving in quantitative (metric) data models. • Wavefront is based on a stream processing approach. It combines and processes millions of data points per second by using its query language and performing complex correlations, precision analysis and long term trend analysis. • It is used when millions of data points ingest per second and manipulate data in real time and providing valuable insight on metric data sets.
2. Hortonworks Data Platform (HDP).	• Hortonworks Data platform is an enterprise which develops and supports open source software and plans to serve big data and associated processing. • Hortonworks Data Platform designed to cater versatile range of processing engines such as Apache Hadoop. It is also integrated with many data management provider to enable their tools and work with them. HDP application involves with both structured and unstructured data. • HDP is used when user wants to integrate and maximize the current security solutions of contemporary data architecture.
3. Amazon Kinesis	• Amazon kinesis is integral part of AWS (amazon web services) for processing big data in real time. High volume (hundreds of terabyte) streaming data from variety of sources (operating logs, financial transactions, social media feeds can be operated and managed by Amazon Kinesis. • Amazon kinesis works through an application (it has its own name, description, version ID, and status), which can be created through an AWS account. This application is controlled and supported by AWS management console or the Kinesis Data Analytics API. These applications are continuously read and process streaming data in real time. • It is used when real time data such as videos, audio, application logs, website clickstreams and many more needs to be analyzed instantly without any wait before processing.

2. DATA ORGANIZATION COMPONENT

After data ingestion, transformed data is deposited (as an accessible repository) for more technical usage, this area is known as Data organization component. There are many ways to arrange ingested data in a Data lake. Most Common techniques for organizing data are **Directory Structure, Version Control, Database management systems, Master Data management and Distributed file systems.** These are explained below.

A. (I) DIRECTORY STRUCTURE

A Directory Structure is a traditional way to keep your files and folders in an organized way and this practice is come from an operating systems' file system. It is basically a hierarchical tree structure. This data organizing technique helps the user to keep track of working folders in the form of catalogs where name and extensions of file display properly and recognizable.

A. (II) VERSION CONTROL

Version control is also known as "file versioning", it has an attribute to save an updated file as a new file instead of overwriting the previous or main source file. Version control is a client-server model and it allows servers to execute different version on different multiple location, even those versions are updating simultaneously.

1.GitHub	• GitHub is a code hosting platform for version control and collaborating working files. It can be accessible by team members working on specific project. • Initially, GitHub server provides online accessible account, in which repository is allocated to organize a single project. Repository contains (synchronized folders, files, images, videos, spreadsheets, data sets many more things which are required for the project. This repository contains branch in which different version of files are residing under repository can be accessible for updating (for all kind of changes)and after committing, it captures the history of all changes so tracking of files is in detail and comprehensive. • GitHub is used when web based graphical interface is required to access on systematic/prototyping files, folders etc. which are keep updating without losing its order and organized file structures.

A. (III) DATABASE MANAGEMENT SYSTEMS (DBMSs).

Database management system is a collection of data records which can be processed and provide useful information. Data inside database management system can be accessed, inserted, modified, managed, controlled and organized to execute various data processing functions. DBMS organizes the data in the form of tables, views, schemas, reports etc. It is mainly capable of handling structured data which can be transformed in to interrelated tables. **ODBC** (open database connectivity) is the driver which is built-in in many DBMS, it allows the databases to integrate with each other or with other applications too. It is most reliable, efficient and secure Data organization component in the Data Lake.

Oracle Database	• Oracle Database is developed by Oracle Corporation and it is relational database management system. Oracle is commonly used to process OLTP(online transaction processing) and DW (Dataware housing) or both workloads. • Oracle operates on basis of instances and these instances rely on background processes. There are basically 9 processes behind oracle instances and they are DBWR(database writer),LGWR (Log writer),SMON(system monitor),PMON(process monitor), CKPT(Checkpoint),ARCH(Archives),RECO(Recoverer),Dnnn(Dispatcher) and Lckn(Log process),these processes are doing functional activities for oracle database. • It is used when you can access required data locally or remotely. Oracle database is containing data sets more than 7 million terabytes.

A. (IV) DISTRIBUTED DATABASE MANAGEMENT SYSTEM (DDBMs)

A distributed DBMS is a set of many logically interrelated databases distributed on the network and it is accessible by authorized users. Distributed database system coordinates with data sets periodically and ensures that modifications and amendments in the datasets made by users or systems are universally updated in the data models. **Apache Cassandra** is an example of DDMS in big data technologies.

A. (V) METADATA MANAGEMENT (MDM)

Metadata management is the most trending solution and it plays a vital role in managing data for organizations in the form of different file extensions and sizes. It introduces policies and processes that ensure that data can be integrated, accessed, shared, linked, analyzed and most important maintained significantly across the organization. Metadata management repositories are allowing data lineage and tracking capabilities, semantic frameworks, analysis features, rules management and metadata ingestion and translation.

B. There are many applications for Data organization component in Data Lake. MongoDB (semi structured data), Apache Hive and Apache Hadoop are most interesting tools of database organization in this category.

1.MongoDB	• MongoDB is an open source, document-oriented database model which provides platform for different form of data. It is non-relational database technology. It comes as NOSQL in big data applications. It is made up of collection of documents. • MongoDB is document based database, in which data is structured in fields and value pairs. This document can be saved in JSON or BSON, .Every entry of data in the document contains system generated unique identifier and it is basic unit of Mongo DB document. And instead of table, in MongoDB "COLLECTION" is the feature comprises of sets of documents and behaves like relational database tables. • It is used due to its flexibility with JavaScript interface which allows MongoDB to query and it supports semi structured data.
2.Apache Hadoop	• Apache Hadoop is an extensive open source frame work, it is well equipped for processing large data sets which come from multiple format and from various sources. The basic core of Apache Hadoop consists of a storage availability and it is known as Hadoop Distributed File system (HDFS), and processing functionality is MapReduce programming model. • Hadoop distributed file system is loaded the data models, it fragments the data information in to different blocks and distribute them in to further nodes in a cluster, than enabling highly efficient parallel processing. This file system replicates each data and distributes it on each individual nodes on different servers, so if you lose the data in any circumstances you can retrieve it from other clusters. The Apache Hadoop architecture is composed of the following models Hadoop common, Hadoop Distributed File system (HDFS), Hadoop Yarn (processing clusters),Hadoop MapReduce(Map reduce algorithm).There are add-on packages available that can be installed with Hadoop suite and these packages contained additional features and functionality as big data functionality (Apache Pig, Apache Hive, Apache HBase, Apache Phoenix, Apache Spark, Apache Zookeeper, Cloudera Impala, Apache Flume, Apache Sqoop,

	Apache Oozie and Apache Storm). • Apache Hadoop is the best practice for batch processing, graph based data processing and dealing with humongous size of data.
3.Apache Hive	• Apache Hive is a data warehouse application that permits fastest accessibility of extracted data in result of SQL like queries from Apache Hadoop (It is an open source framework for containing huge amount of datasets). • Apache Hive interprets the input program, code in HiveQL (SQL-like) language to one or more Java MapReduce, Tez, or Spark jobs.(these all are execution engines in Hadoop Yarn).Apache Hive arranges the data in to tables for the Hadoop Distributed File System) and execute jobs on a cluster to generate the required reports. • It is used because of its utility of SQL which is more direct and easy for database professionals to practice and Apache Hive is mainly generating an analytical reports for data analyst.

3. DATA SECURITY AND GOVERNANCE COMPONENT

When the capacity of storage of data is keep increasing than it is required to pay more attention and take more control on data access by introducing data governance policies. The distribution of files related to any project or application are scattered on cloud and on different networks and if there is no specific restriction on accessibility, this can be lead to further risk of data exposures. There are several issues which are required governing policies for access of right data and proper authority to define and modify the data.

Duplication of data files is very common and poor practice and it exists on very large level across an enterprise internal network or/and on cloud accessible data files. Duplication of data files directly effects the storage capacity, the issue requires proper discipline to follow not to duplicate data and also not to store same working or data files with **multiple versions**. If there is no governance on right of data access, users can work from any personal workstation and after modification ,when files are transferring back to the network or cloud,

these files are highly outside the protection of authentic servers, and also not containing the versioning and audit controls of the enterprise management system.
Multiple version of the same file are moved around in the same application can cause conflict information. Without access permission, users are **sharing data files**, which are holding sensitive information. These are the glitches in the data environment which are required governing policies which are strictly impose on data access rights and data definition and modifying rights.
Data governance is a centralized control and monitor mechanism to manage integrity, usability and data availability. Data governance performs functions such as setting data management parameters, resolve data duplication and redundancy and many more to secure usage of data in Data Lake.

1. IBM Data Governance	• The IBM data Governance was formed in 2004 to explore big data challenges and develop solutions. • It works with structured and unstructured data, it has feature like flexible data governance strategy, data cataloging, and obtaining useful information for big data projects. • It is usable when data sets are structured and unstructured and users are looking for data sets proper information (name, location etc.)
2. Cloudera Navigator	• Cloudera Navigator is the Governance solutions proposed by Cloudera Enterprise. It is a search based interface. • Cloudera works for Apache Hadoop is most integral part of Data Lake in big data technology. It main functions are data discovery, continuous optimization, policy enforcement, Meta data management and many more. • It is useful when in database management component is having Apache Hadoop.
3. Acaveo	• Acaveo is a client based software partner with Microsoft for storage and information governance. • It works with large volumes of unstructured data.it has ability to analyze files, data management, data classification, migration and many more.

	• This governance tool is used when very huge amount of unstructured data is generating, accessing, modifying, and inserting regularly in data models of Data Lake.

B. In big data, it is required to understand that involvement of very large data from complex structured and unstructured data can be protected otherwise non authorized accessibility of data can create new relations, combine different data sources and non-recognizable user can retrieve the information of data. There are sets of risk areas that include information life cycle, the data creation and collection process and the lack of security procedures.
CSA (Cloud Secure Alliance is a nonprofit organization and supports big data security and privacy challenges. According to CSA there are for main existing challenges in security, trust and privacy in big data. These are mentioned below.
(I) Infrastructure Security.
(II)Data Privacy
(III)Data management and Integrity
(IV)Reactive Security
There are issues of security and privacy with big data devices, which suppliers need to understand and avoid them. Some are mention here: Insecure web interface, Insufficient Authentication/Authorization, Insecure Network Services, Lack of Transport Encryption, Insecure Cloud Interface and so on.

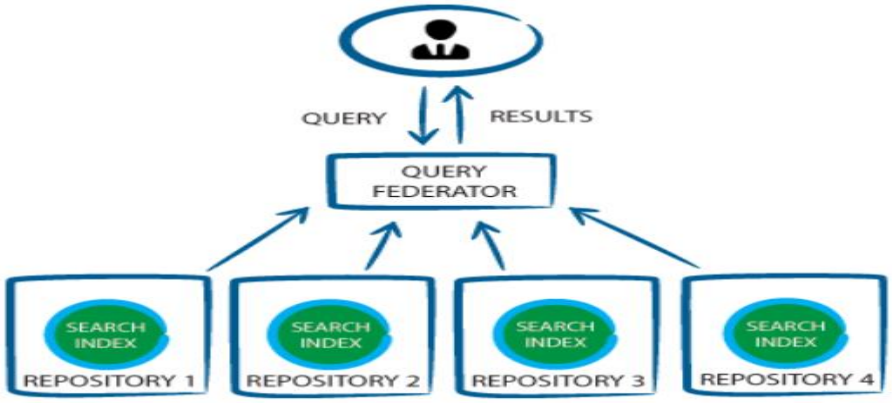
1.Qualys 2.Gemalto	• Qualys provides solutions for cloud security, serves as SaaS model. • It works as a cloud based solution and finds vulnerabilities on all networked assets, servers, network devices and workstations. It can access any device that has an IP address. • It is used when precise network monitoring and security is required. • Gemalto is one of the competent enterprise dealing with digital security. • Gemalto SafeNet provides solution for data encryption and tokenization to enhance security levels in big data technologies. Its framework allows integration with MongoDB, Cloudera, Couchbase, DataStax, Hortonworks, IBM and Zettaset.
	• This is used when intranet security is required at a very efficient level.

4. INDEXING AND SEARCH COMPONENT

Federated search is a technology that permits simultaneous search of multiple searchable resources and retrieve information. Federated search requires centralized correspondence of the searchable resources, coordination of the queries transmitted to the individual search engines and provides search result returned by them. Index-time merging, Query-time merging are both ways of performing Federated search. Hybrid Federated search is the combination of index-timing and Query timing search.

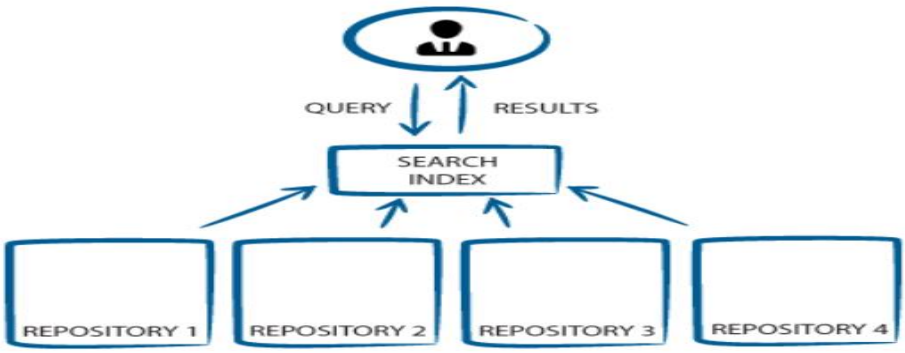
A. (I) QUERY-TIME MERGING

A query federation grabs the query and sends it to many search engines, then federation waits for the response from search engines, after receiving the results, it updates them immediately. This approach relies on repository of data to provide a search function. Google search engine is the appropriate example for query-time merging.



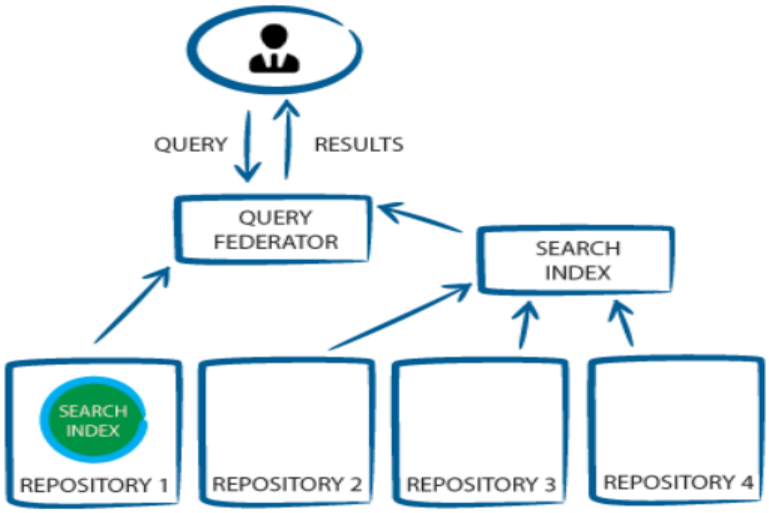
A. (II) INDEX-TIME MERGING

Index-time merging is inspire from typical traditional search procedure, central index is the core component and it captures content and supplies the results with the help of relevant algorithms.



A. (III) HYBRID FEDERATED SEARCH

The precise solution is a hybrid approach, in which content is indexed centrally and data repositories are federated to query time.



1.SerachBlox	• SearchBlox is well known enterprise search engine which deals with sentiment analysis and text analytics solutions. • It performs normal/regular searches as well as multifaceted searches with the approach of index up to date information on data content and deliver pertinent search results. It involves in query syntax, synonyms, auto suggest servlet and highlighter servlet. • It is used because of its diversity of handling public and private both searches. It searches from websites, file systems, databases, cloud environment, web portals and custom generated contents.

B. There is wide variety of tool and technology for indexing and searching in bid data. ElasticSearch, Amazon Cloudsearch, Graylog2, Algolia, Apache Solr, Ambar and many more.

1.ElasticSearch	• ElasticSearch is a search engine based on Lucene library. It provides full text search engine with web interface and schema free JSON documents. It is developed in java and supported by almost all clients such as java, .NET, Python and other programing languages. • Elasticsearch works with the functionality of inverted index. Inverted index is the core of Elasticsearch and makes it different from others such as MongoDB, Cassandra and so on. • It is most flexible and precise approach in searching and indexing and supported by many programming languages.
2.Apache Solr	• Apache Solr is an open source search platform based on java library called Lucene. • Solr works with HTTP and XML.APIs programs provide connectivity with JSON, Python and Ruby. Like Elasticsearch Apache Solr is also created inverted index and completes its searches. • It can be used because it is popular search practice for web sites.
3.Algolia	• Algolia is a hosted search engine capable of delivering real time results. • Algolia works on its powerful APIs, these APIs allow Algolia to search quickly and effortless. • It can be used because it is fastest and provides search results in 100ms around the world.
4.Ambar	• Ambar is an open source document search engine with automated crawling, OCR and instant full text search. • Ambar works in a unique way, which is to deploy Ambar with a single docker-compose file. Ambar supports all popular document formats, execute OCR, tags the documents and use simple REST API to integrate Ambar itself to integrate Ambar into your workflow.

	• It is used when full text document search is required through files.

5. ANALYTICAL COMPONENT

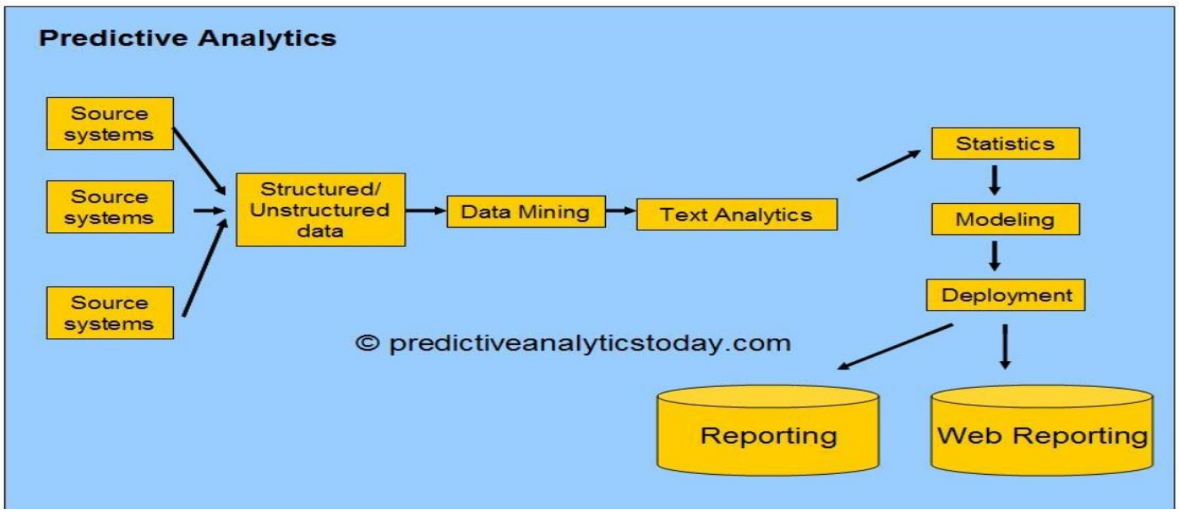
Big data is only useful in the situation when it is leveraged to drive decision making. This unique accessibility in to diverse data delivers meaningful insights. Process of extracting insights can be divided in to five stages.

A. (I) TEXT ANALYTICS

Text analytics is a technique that extracts valuable or required information from textual data. Emails, blogs, survey responses, social network feeds, call center logs and many more example of textual data held by enterprises. Text analytics plays a vital role in statistical analysis, machine learning and computational linguistics. It allows large volume of data to day generated text in to meaningful summaries, which promotes evidence based decision making. There are ways to apply text analytics such as IE (information extraction) technique which extracts structured data from unstructured text with the help of IE algorithms. Another is Text summarization techniques which has functionality to provide sump summary of large single document or multiple documents.

A. (II) PREDICTIVE ANALYTICS

Predictive analytics use previous data information to forecast or predict the future. There are companies using this analytical approach for sales and scoring. As analytics are growing day by day, some enterprises are using and practicing predictive analytics for the entire sales process, analyzing lead source, number of communications, social media, CRM data etc. Predictive analysis are eligible to perform complex forecasts. Predictive analytics interpret big data for the benefits of the business. Predictive analytics apply on both structured data (age, income, sales, gender etc.) and unstructured data such as textual data (social media content, open text etc.)Which need to be extracted from the text, along with sentiment, and then used in the model building process.



1.SAP Predictive Analytics	• SAP Predictive Analytics is a business intelligence software and designed to organize the large datasets and predict future outcomes and better understanding of uncover hidden risks. • SAP Predictive Analytics composed with two products in one interface, one is automated analytics for accessing insight and another is Expert analytics for predictive analytics, it connects with most of the environments including (unstructured, semi structured and structured data models). • SAP Predictive analytics provide automated processes for business users and data analysts to generate predictive models. (It is used when users are dependent on automated process).
2.SAS Predictive Analytics	• SAS statistical analysis system is a software suit developed by SAS institute for multivariate analysis, data management, business intelligence and Predictive analytics. • SAS programs composed of two steps, one is DATA steps, which extract and manipulate data sets, and PROC steps, which analyze the data (series of statements are occurring for execution of each steps). • It can be used when user friendly environment is the priority, because SAS provides a graphical point and click user interface for non-technical users and SAS suite can do data manipulation from variety of sources and execute statistical analysis.
3.IBM Predictive Analytics	• IBM predictive analytics is a flexible predictive analytics suite designed serve all analytical processes, data collection, reporting and many more. • It works with support of its robust SPSS Analytic server built to function as predictive analytics for big data. The module works simultaneously with the SPSS Modeler to deliver an integrated platform to Spark applications and Hadoop distributions to utilize their data sets. It performs real time processing for deep analysis, machine learning while consuming minimum code and fastest results.

	• It is used when intuitive solutions are required because it offers comprehensive predictive analytics capabilities and extensive collection of algorithms on a single platform.

A. (III) DESCRIPTIVE ANALYTICS

Descriptive analytics is the interpretation of historical data to comprehend all the changes that come orderly in data information of particular business sector. It works with financial metrics to compare with time to time changes in raw data for example sales growth, revenue generation and many more. Descriptive Analytics take raw data and resolve that data to draw conclusions that are useful for the users.

A. (IV) PRESCRIPTIVE ANALYTICS

Prescriptive analytics is related to both descriptive and predictive analytics. As descriptive analytics promote historical insight that what happened from different time frame and predictive analytics provide forecasting on data sets that what will happen. It illustrates the implication of each decision option. It improves the accuracy of predictive analytics. The effect of the prescriptive analytics also depends on the appropriate decision models and how they impact on analyzed decisions.

A. (V) DIAGNOSTIC ANALYTICS

Diagnostic analytics is an advanced analytics which inspects data by drilling down, data mining, looking for correlations and many more. It is deeper analytical approach towards data and understand the causes of events and behaviors.

B. There are many tool and technologies for analyzing the big data. SAS Tools, Microsoft ML Platform, Amazon Platform and Apache Mahout are some very famous and popular examples of data analytics software.

1.Apache Mahout	• Apache Mahout is a library of scalable machine learning algorithms, which is part of Apache Hadoop and MapReduce paradigm. • It works as a data science tool to function on data which is stored in Hadoop Distributed File System and with the help of automated process it finds the meaningful patterns in those big data sets. • It can be used because if supports collaborative filtering, Clustering, Classification and frequent itemset mining.

.

2.RapidMiner	• RapidMiner is a platform which serves in text mining, machine learning, Dara preparation and predictive analytics. • It works as a client server model and serves on premises or in public or private cloud infrastructure. RapidMiner is written in java programming language., it provides GUI to design and process workflows .APIs are also embedded in RapidMiner, it has flexibility to add on more plugins which are already created or can be created by developers on RapidMiner Market place. • It can be used when an advanced analytical solution is required through built-in template based frameworks and reduce errors by almost eliminating the need of writing code.

6. VISUALIZATION COMPONENT

A. Visualization of data is most favorable practice to make users comprehend visually of type and information of data carrying in the data sets. There are many simple technique to visualize data in general level. Number charts, Maps, Pie Charts, Gauge Charts, Timeline, Tree and many more are the example of data visualization. Excel is one of the technique or tool for data visualization.

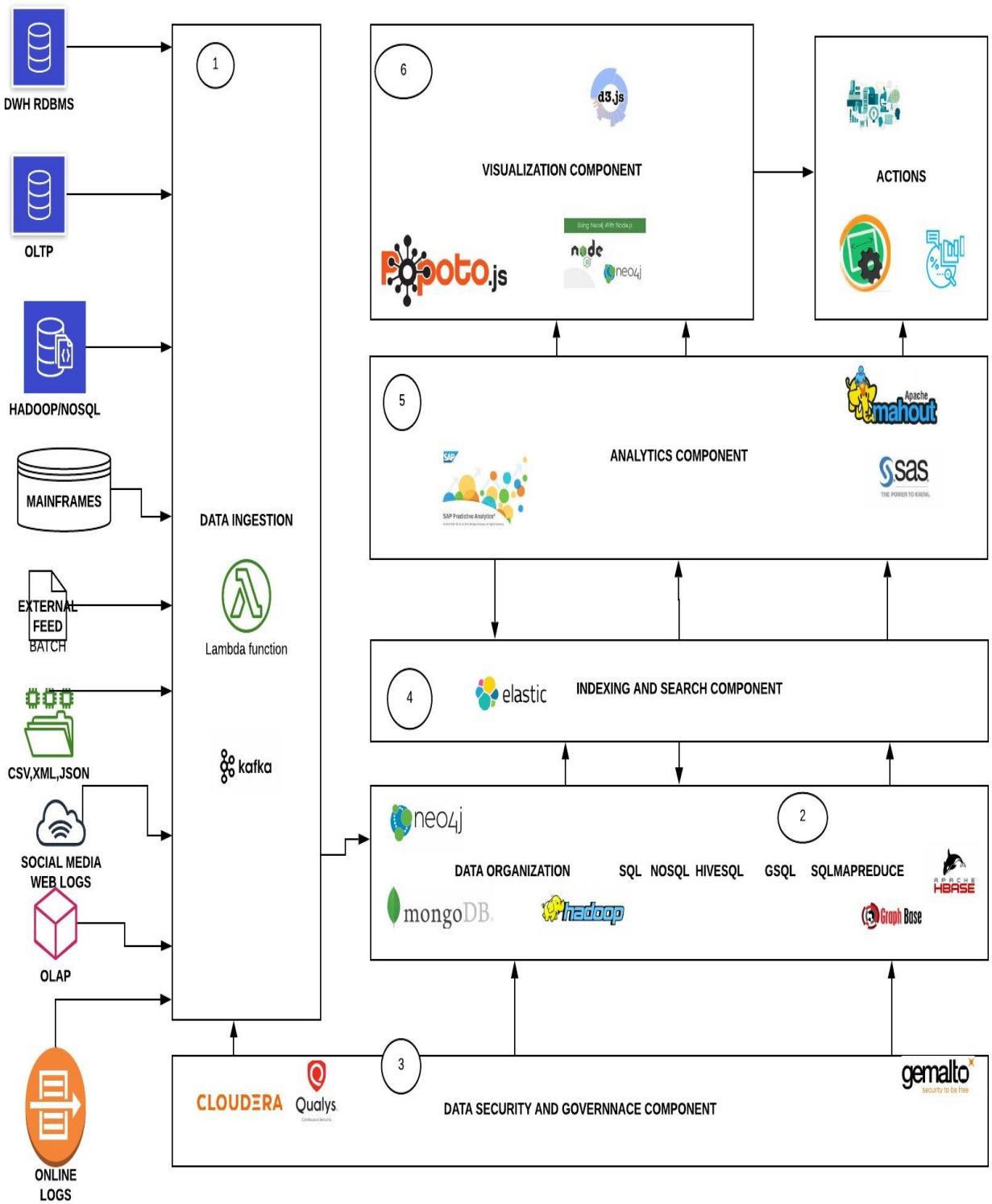
1.Microsoft Excel	• Excel is a commercial spread sheet application produced by Microsoft software. It has features in which by performing simple calculation with data set, visualization of that estimated data is possible using graphing tools. Data can display as charts, histograms and line graph. • User can arrange data on spread sheet of excel to view various possibilities by calculation data sets or by visualizing them. Visual basic is used in Excel, it permits user to create a variety of complex numerical method. Programmers are even have option to code using VB Editor, including windows for writing code, debugging and code module organization. • It is used when u have small data sets (not like big data), there is no requirement of interactive data visualization and simple graphs are needed.

B. Big Data visualization promotes the presentation of data in the form of graphical structure and definitely this approach makes it easy to understand and interpret by professionals or even common users. But in big data the category of visualization is not only stick with simple line, bar and pie charts, there are advanced level of graphical presentation is available such as heat maps, fever charts, enabling decision markers to investigate data sets and find out correlations or unexpected patterns. Big Data visualization depends on powerful systems to ingest data from different sources and transform it to generate graphical representation that allow people to comprehend very large amount of data in seconds. There are many example of Big Data visualization in big data such as Linear, 2D/Planar/geospatial (cartograms, Dot Distribution maps, Proportional symbol maps, Contour maps), 3D/Volumetric, Temporal (Timelines, time series charts, connected scatter plot, arc diagrams) and Multidimensional (Pie charts, histogram, tag clouds, bar charts, tree maps, heat maps, spider charts),Tree/Hierarchical. In big data technologies there are many progressive tool available on different platforms which provide interactive and multidimensional visualization of data.

1.Microsoft Power BI	• Microsoft Power BI is a business analytical tool and provides visual analysis on data also to enhance strategies performance in different business units. • Power BI is a collection of software services, apps and connectors that work together to bring different sources of data on one place and visually analyze it and try to comprehend the interactive insights. • It can be worked when datasets are importing from Excel, cloud based repositories or even hybrid data warehouses and delivers appropriate reports and visualization.
2.Tableau	• Tableau software is one of the most favorable data visualization tool. It changes the raw data in to an easily understandable format without any coding knowledge. • Tableau engine connects to the many different platform and access data sets. It creates dashboards and share with users in the form of static file. The user can access dashboard reports by using Tableau Reader. Dashboard reports can be extracted from Tableau Desktop and publish on Tableau server. • It can be used when queries can be translated in to visualization. It manages metadata, on-technical users can work in Tableau because it has facility to create no code data queries.

3.Sisense	• Sisense is a BI platform. It does an ad-hoc analysis of very high volume of data and connects with all the sources of data and gathers information and manages entire BI workforce. • Sisense has in-Chip engine makes data accessibility fast. It generates various queries and get faster results. Sisense is adaptable to commodity server infrastructure and no need to install high end servers. It is facelifted with single –stack system which allows multiple task and data integration. It has drag and drop interface which makes it simple to combine with big data sets. It has variety of visualization options. • It can be used when instant deployment is required with the minimum total cost of owner ship. It provides end to end BI solutions
4.D3.JS	• D3 is a JavaScript library and frame work creating visualization. It generates visualization by combining data and graphical elements to the Document Object Model. It allows direct changes in DOM which is unique feature from other visualization toolkits. • D3 is JavaScript frame work and it uses SVG (scalable vector graphics) to generate graphical elements. It also uses CSS and chaining syntax. • It can be used when Animation and display of dynamic data is required. It cannot be used when statistical analysis is required.

Part 2. Data Lake Architecture



https://www.lucidchart.com/invitations/accept/33af9808-2874-45bb-ab08-6473ead3007d

AISHA HASSAN SHAH **45920842**

DETAILS OF GRAPH BASED SEARCH DATA LAKE

1. DATA INGESTION COMPONENT

- In this Data Lake, Lambda and Kappa architecture both techniques are used.
- Lambda architecture supports real, batch and stream processing, in which it easily consumes all graph, maps and image related data files.
- It extracts data from different sources of data available on various platforms and transform it for further accessibility.
- It is used when, source data is available with all kind of file extensions and formats.

2. DATA ORGANIZATION COMPONENT

- In this data lake, Apache Hadoop and Neo4j are main technologies used for NOSQL and other different types of data sources.
- These both technologies support graphs and images and extract this variety of data in to the system really well.Neo4j is a graph data base in which data points are connected to each other and those points are called nodes and they have directional properties.
- It is used to facilitate graph presence in source data.

3. DATA SECURITY AND GOVERNANCE COMPONENT

- This data lake has main objective is graph search so the security and governance is also Participates to take care of all aspects of graph search. Cloudera Navigator is the governance and security software which suits best on this requirement.
- It is collaborated with Apache Hadoop and Apache Hadoop deals with graph.
- It is used to provide governance and security of all files, transportation, extraction etc.

4. INDEXING AND SEARCH COMPONENT

- For indexing and search in this data lake, Elastic Search is implemented to support all kind of search related to graphs or image oriented data bases.
- It works with the strategy of inverted index that helps in indexing graphs too.
- It can be used with all kind of data sources, it searches for all.

5. ANALYTICAL COMPONENT

- For analytics purposes, Apache Mahout is available and promotes graph search.
- It has very strong integration with Hadoop file systems and analyze graph data sources really well. It has automated processes to search and locate graph nodes.
- It is used when, image analysis is required.

6. VISUALIZATION COMPONENT

- D3 is a JavaScript library and frame work creating visualization.
- D3 is JavaScript frame work and it uses SVG (scalable vector graphics) to generate graphical elements. It also uses CSS and chaining syntax.
- It can be used when Animation and display of dynamic data is required. It cannot be used when statistical analysis is required.

WORKFLOW OF GRAPH BASED SEARCH DATA LAKE

- Data is entered in to the system from many different data sources in to the data ingestion component.
- After proper transformation, data sets are loaded in to the data base component in to their respective database management systems.

- Data security and governance is taking care of all components and try to avoid any kind of violation of rights, authenticity and authorization.

- Indexing and search component is linked with data base component and Analytics component, it is two way traffic between them.
- Analytics component is extracting required data from indexing and search component and generating reports.

- Visual component is extracting information data from analytics component and providing outcome to the external users/Actors.

REFERENCES:

https://www.oreilly.com/library/view/the-enterprise-big/9781491931547/ch01.html

https://www.datamation.com/big-data/structured-vs-unstructured-data.html

https://www.codecademy.com/articles/what-is-rdbms-sql

https://www.datamation.com/big-data/structured-vs-unstructured-data.html

https://www.geeksforgeeks.org/difference-between-structured-semi-structured-and-unstructured-data/

https://www.predictiveanalyticstoday.com/data-ingestion-tools/

https://aws.amazon.com/kinesis/

https://docs.aws.amazon.com/kinesisanalytics/latest/dev/how-it-works.html

https://www.sciencedirect.com/topics/computer-science/distributed-database-management-system

https://guides.github.com/activities/hello-world/

https://instr.iastate.libguides.com/dmp/step2

https://www.quora.com/How-does-oracle-database-work

https://www.sciencedirect.com/topics/computer-science/apache-hadoop

https://www.computerweekly.com/feature/Data-governance-the-importance-of-getting-it-right

https://www.researchgate.net/publication/281404634_Security_and_Privacy_Issues_of_Big_Data

https://www.softwaretestinghelp.com/data-governance-tools/

https://www.researchgate.net/publication/281404634_Security_and_Privacy_Issues_of_Big_Data

https://www.searchtechnologies.com/blog/federated-search

https://www.lucidchart.com/invitations/accept/33af9808-2874-45bb-ab08-6473ead3007d

https://www.cloudtp.com/doppler/how-to-guide-architecture-patterns-to-consider-when-designing-an-enterprise-data-lake/

https://www.capgemini.com/2014/08/you-have-to-manage-your-data-lake-the-fallacy-of-technology-being-magic/

https://www.slideshare.net/Pivotal/10-thingsdatalake-pivotalstratanyfinal
https://www.sqlchick.com/entries/2016/7/31/data-lake-use-cases-and-planning

https://neo4j.com/blog/time-for-single-property-graph-query-language/?ref=blog
https://neo4j.com/blog/why-database-query-language-matters/

https://www.sciencedirect.com/topics/computer-science/apache-hadoop