COMP8210 Big Data Technologies

ASSIGNMENT 1

STUDENT NAME: MOHAMMED SADIQ ABUWALA

STUDENT ID: 45921407

Data Ingestion Component

(A)

Structured data:

It is often categorized as qualitative data. Structured data is often stored in data warehouses and the data itself is highly organized. Such data exists in predefined formats. It is more commonly utilized by business users because structured data does not require an in-depth understanding of different types of data and how they function. With a basic understanding of the topic relative to the data, users can easily access and interpret the data (). However, flexibility and usability can be limited when using structured data. This is because such data has a predefined structure and hence can only be used for its intended purpose. Finally, changes or updates in data requirements can be time and resource intensive when using structured data because they are usually stored in data warehouses which possess rigid schemas. SQL is used as the programming language for structured data.

Unstructured data:

It is often categorized as qualitative data. Unstructured data is not predefined and therefore best managed by non relational (NoSQL) databases. Unstructured data is stored in its native format. Some examples of unstructured data could be: social media activity, surveillance imagery, etc.

Semi structured data:

A type of data that does not entirely match the description of structured data but utilizes markers or tagging mechanisms as such that allow searching and separation of semantic elements. Semi structured data is often found in object oriented databases. Typical examples of semi structured data include the XML and JSON file types.

Batch

In batch processing, we wait for a certain amount of raw data to “pile up” before running a job (). This means that the data used for the job may be anywhere between an hour to a few days old. Batch jobs are usually run on a set schedule or in some cases when the amount of data collected reaches a certain threshold.

Micro-batch

In such a method, batch processes are run on small accumulations of data. This means that data could be available in near real time. It is useful when fresh data may be required but not necessarily in real time. An example of a scenario when micro batch processing may be useful is in web analytics (i.e. analysing clicks on a page).

Real time

In a real time job, we process data as soon as it arrives within the storage layer. Real time processing could be used when it is important to analyze or serve data closest to its arrival. Examples of such scenarios are real time advertising and fraud detection.

(B)

Apache kafka: It is an open source, low latency, unified, and high throughput platform for managing real time feeds. By utilizing Kafka, enterprises can utilize a single cluster to be the backbone for their organization.

Sqoop: A popular ingestion tool where data from RDBMS is imported into Hadoop. Sqoop runs on the MapReduce framework.

Azure Data Factory: It is a fully managed and serverless data integration service.

Wavefront: Hosted platform for data ingestion,storage, reporting and visualization of data.

Data Organization Component

(A)

Data Version Control: An open source framework that allows reproducibility of data within machine learning environments. The framework consists of various tools that allow the tracking of changing versions of data. Therefore, bu utilizing data version control, the user does not have to manually track which model was trained an upon which version of data.

Document Management Systems: Generally relates to the usage of a software to store, organize, process, and track data efficiently. This data may refer to input of various types such as PDFs, images, and word processing files.

(B)

SQLite: Relational database engine that is severless, self contained and requires zero configuration. SQLite is easy to set up and does not require too many configurations in comparison to MySQL.

MySQL: it is also a relational database management system based on SQL. While SQLite is severless, MySQL follows a client and server architecture and thus requires a server to run (MySQL, 2021). MySQL can more efficiently handle bigger databases. They can be more suited towards users who require multiple user access, strong security such as authentication features, and apps that require large databases.

PostgreSQL: Supports high-level programming languages such as Python, Java, etc. Furthermore, it supports SQL as well as JSON querying. It supports more advanced data types which are not supported by MySQL.

MongoDB:

It is open source and available to any vendor (i.e. vendor agnostic). MongoDB is a document oriented NoSQL database. It makes use of collections and documents instead of traditional relational databases which utilize table and rows (What Is MongoDB?, 2021). In MongoDB, these documents consists of key-value pairs and these are the basic units of data.

DynamoDB:

DynamoDB also utilizes a key-value and document database like MongoDB. However, DynamoDB is a fully managed AWS service and hence only available on AWS. Furthermore, DynamoDB supports fewer data types and smaller item sizes in comparison to MongoDB (Comparing DynamoDB and MongoDB, 2021). Finally, DynamoDB structures its databases similarly to relational databases. It utilizes tables that contain items which in turn have attributes. In comparison, MongoDB stores its information in a JSON like format in which the main stored objects are called documents and these are subsequently grouped into collections.

Azure Cosmos DB: It is Microsoft's globally distributed, and proprietary database. It is generally classified as a NoSQL database. Furthermore, it is horizontally scalable and schema agnostic.

DATA SECURITY & GOVERNANCE COMPONENT

(A)

In the following section we provide a set of requirements for governing the right data access and the rights for defining and modifying data:

- Fine-grained access control (FGAC): Since data processed by Big Data analytics platforms often refer to user personal characteristics, it is important that access control rules can be bound to data at the finest granularity levels. For the protection of sensitive or personal data, FGAC is widely recognized as a fundamental component.
- Context Management: Access control to data should also be limited based on context based constraints such as geographical location or specific time periods. Hence, in these cases, access is granted when certain conditions based on the property of the environment are met.
- Efficiency of access control: The characteristics of the Big Data scenario, such as the distributed nature of the considered platforms, the complexity of the queries, and the focus on performance, require access control enforcement strategies that do not compromise the usability of the hosting analytic frameworks (). For example, in traditional relational DBMS, Fine Grained Access Control (FGAC) is achieved via two main approaches:
    - View based: Users are only allowed a view of the dataset based on specified access control restrictions.
    - Query rewriting: In such a process, the query is modified at runtime by injecting restrictions instead of precomputing the authorized views.

(B)

Azure Active Directory (Azure AD):

It is a cloud based user management system designed for all Azure web applications and infrastructure on the cloud (Azure AD Connect: Seamless Single Sign-On, 2021). Therefore, it is highly tailored for Windows based infrastructures. Most organizations utilize an Active Directory instance on premise and then combine them with another Microsoft solution called Azure AD Connect to connect to Azure AD.

Okta:

It is a SaaS and cloud based identity and access management system. Some examples of the identity stores that it can connect to include AD, LDAP and also several third party identity providers. Okta connects everything from cloud to ground with 1400+ SAML and OpenID Connect integrations, password vaulting, RADIUS and LDAP support, and connections to third-party legacy SSO solutions (Single Sign-On, 2021). Okta is also able to provide context based two factor authentication.

OneLogin:

It is platform agnostic and has a similar feature set in relation to Okta:

- Passwordless authentication: In such an authentication method, a user's device password acts as the first authentication factor. In the second step, a OneLogin certificate that is preinstalled on the device is verified (OneLogin vs. Okta: The Difference, 2021). Hence, the user is not required to input any password at the time of authentication and thus it is called Passwordless authentication. Okta also features passwordless authentication via implementations such as mail based magic links and smart cards.

- Reports: OneLogin offers detailed reporting. It generates four types of reports (i.e. users, apps, events, and logins). In comparison, Okta security reports are categorized into three sections: user activity, security and system log queries.

- SmartFactor Authentication: Block high risk logins based on a calculated risk score that is determined during each login attempt. Okta utilizes ThreatInsight. It stores information about malicious or suspicious IP addresses to thwart potential cyber attacks such as Phishing, Brute Force Attacks, etc.

- Shared logins: It is an important feature that OneLogin offers. It allows a user to share access to applications.

INDEXING & SEARCH COMPONENT

(A)

Federated search allows a user to search through many different data sources at once via a single query. In such a process, the federator collects results from multiple search engines and presents them to the user in a single user interface. This makes the content more useful, and discoverable. This is because by being able to search from a single location, the user is able to find the content with fewer page views or clicks. This also means that less time is spent getting to the data and hence a better search experience and efficiency is achieved. An example of federated search is the Windows search or the MacOS Spotlight. It returns all kinds of search results such as documents, apps, media, and more in a single user interface.

(B)

Elasticsearch

It is a JSON based RESTful search and analytics engine based on the Lucene library. It allows the user to perform and combine various types of searches such as structured, unstructured, geographical and metric data (What is Elasticsearch?, 2021). Furthermore, it provides scalable and near real time search for large volumes of data. As Elasticsearch utilizes a powerful aggregation module, it can also be useful as an analytics engine.

Apache Solr

It is an open source server for performing search and also based on the Lucene library. Apache Solr provides its search capabilities via HTTP requests. Furthermore, it offers distributed and full text search, NoSQL features and near real-time indexing. Finally, it can be integrated with popular big data tools such as Hadoop and Spark. Apache Solr can handle ingest data from various file types such as XML, CSV, databases, PDFs. and Microsoft Word documents. With additional native support for the Apache Tika library, it is further able to offer compatibility for over one thousand file types. Apache Solr performs best for static data while Elasticsearch is more suited towards rapidly changing data (Solr vs. Elasticsearch: Who’s The Leading Open Source Search Engine?, 2021).

ANALYTICS COMPONENT

(A)

Math based techniques:

Time series analysis: Measurements here are collected over predefined periods of time. Thus a collection of organized data known as time series is produced.

Regression analysis: A process for sorting out variables that indeed have an impact. It is a statistical process for estimating the relationship between a dependent variable (i.e. outcome) and one or more independent variables (i.e. features). Some examples of regression analysis include Linear and Ridge Regression.

Machine learning based techniques:

Artificial Neural Networks (ANN): It is a system that allows its structure to adapt based upon the information that flows through the network. ANN are known to be highly accurate and can accept noisy data. As they are highly dependable, they are more widely suited towards use cases such as forecasting applications where a high level of accuracy is required.

Decision Trees: In such a technique, a tree structure is generated based on regression or classification models. The dataset is broken down into smaller subsets while its associated decision tree is also generated. As decision trees break down complex data into more manageable parts, they are very useful in machine learning and data analytics. Some of its popular use cases include prediction analysis and data classification.

Techniques based on visualization:

Some examples of data analysis techniques in this area are Bar charts, Pie charts, Gantt charts, etc.

(B)

Azure ML: An advanced analytics platform that provides end to end services and is cloud based. By using AzureML, users can import data, build, train, deploy models, and predict outcomes via a web browser. All this is achievable via a drag and drop interface (Azure Machine Learning Service - Part 1: An Introduction, 2021). Once a model is

developed, it can easily be deployed in the form of a web service hosted on the Azure platform.

Amazon SageMaker: It was released in 2021 as a subsequent evolution to Amazon ML. It hosts Jupyter Notebook to explore and visualise data and thus a user is required to have a high knowledge of programming / data engineering. In contrast, the Azure ML's drag and drop interface caters to a larger user base. Amazon SageMaker can scale models when required and also clean/tune the data automatically for best accuracy. Models are easy to deploy as they are hosted on an auto scaling cluster of Amazon EC2 instances (What Is Amazon SageMaker? - Amazon SageMaker, 2021). For the purpose of high performance and availability, these Amazon EC2 instances are spread throughout many zones.

SAS Visual Text Analytics:

A comprehensive solution to identify and categorize text data (SAS Visual Text Analytics, 2021). It offers a large variety of modeling approaches to get the most value from unstructured data. Insights are gained by combining linguistic rules, natural language processing and machine learning.

VISUALIZATION COMPONENT

(A)

Line Plot: It is one of the simplest techniques which consist of plotting the relationship of one variable over another.

Bar charts: Quantitative analysis of different groups or categories. In this case, categories are represented via bars which may be vertical or horizontal based. The height or length of each bar represents the value of the category.

Scatter plots: Used to examine the correlations between two variables.

Box and Whisker Plot: Helps with the representation of large data distributions and to also detect outliers. Essentially, the plot displays five statistics: minimum, lower quartile, median, upper quartile, and maximum. Therefore, it allows a user to see a summary of the distribution of data.

Network diagrams: A visualization technique that can be utilized for semi-structured or unstructured data. In such a diagram, individual actors are represented as nodes. For example, network diagrams can be used in the analysis of social networks.

Correlation matrix: Depicts the correlation between all the possible variables in a dataset. They help in summarizing, identifying and visualizing patterns in a large dataset. Correlation matrix is also used in conjunction during other statistical analysis techniques such as regression models.

.

(B)

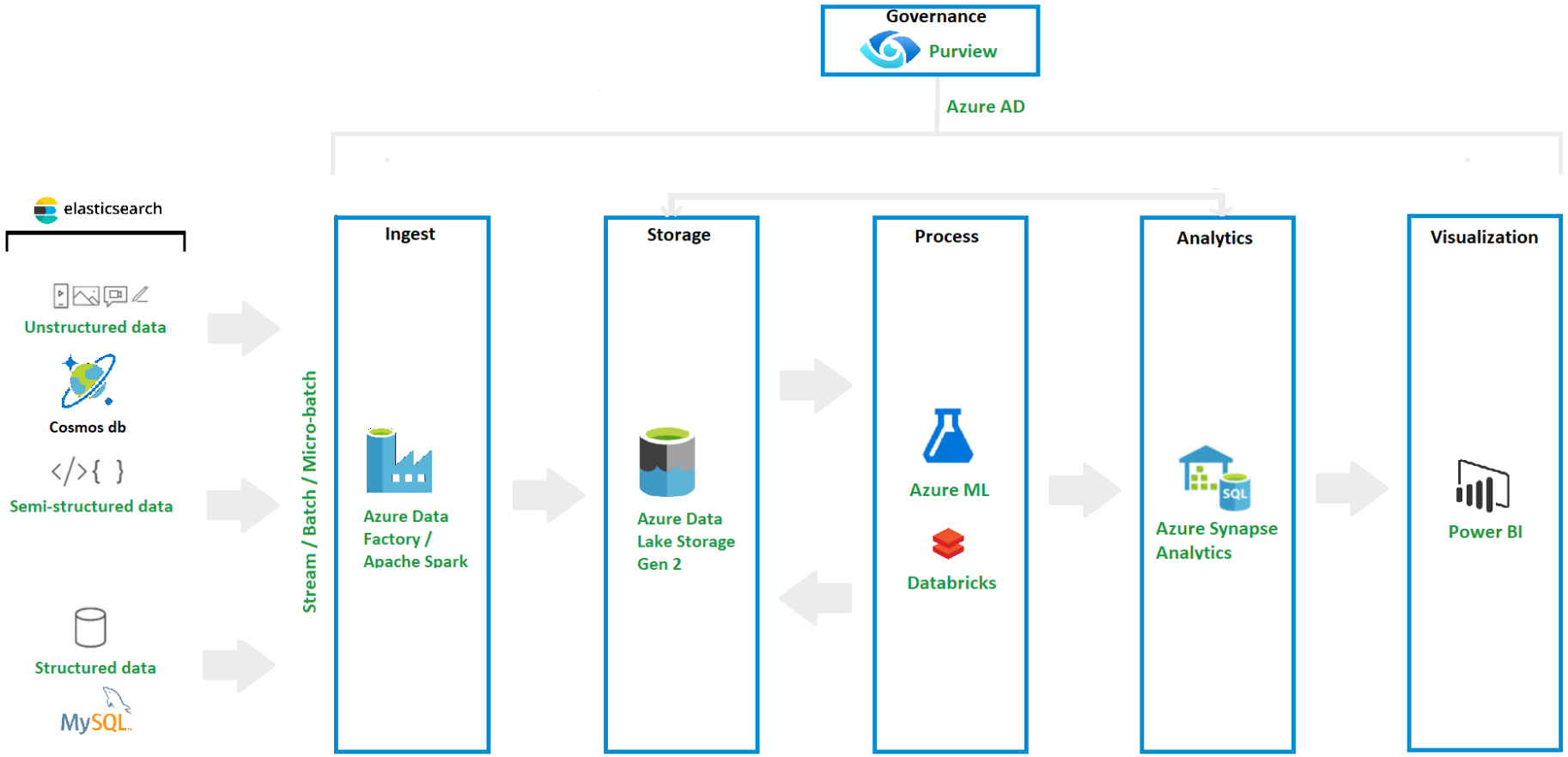
PowerBI: A business analytical solution by Microsoft that lets the user visualize data and share insights across the organization (What is Power BI, 2021).

SAS Visual Analytics: Offers a single user interface for data exploration, reporting and analytics.

D3.JS: It is a Javascript library for manipulating documents to produce various types of charts via web standards such as HTML, Canvas, and SVG.

Student Name: Mohammed Sadiq Abuwala
Student ID: 45921407

PART B - DATA LAKE ARCHITECTURE



Student Name: Mohammed Sadiq Abuwala
Student ID: 45921407

DESCRIPTION OF PROPOSED DESIGN PATTERN:

DATA LAKE ARCHITECTURE: The proposed data lake design pattern is an Azure based data lake architecture.

DATA INGESTION COMPONENT: This tier depicts the process of source data being loaded into the data lake. Data ingestion can occur in real time, batch and micro-batches. This design pattern takes structured, semi structured and unstructured data as input. Structured data may be in the form of relational tables and are stored in MySQL. Semi structured and unstructured data is stored in Cosmos DB. Cosmos DB is a fully managed NoSQL database offered via the Azure platform. Data ingestion is conducted via the Azure Data Factory. It allows real time, batch, and micro batch ingestion.

DATA STORAGE COMPONENT: This tier relates to the storage of ingested and transformed data. An Azure Data Factory has been utilized in this design pattern as it is an Azure based data lake. It supports HDFS semantics and also works with the Apache Hadoop ecosystem.

DATA PROCESS COMPONENT: This section relates to the organization, preparation or transformation of data. Data transformation is conducted via the Azure and Apache Spark powered Databricks. Furthermore, Azure ML is utilized in order to build, train, and test models.

DATA SECURITY & GOVERNANCE: This tier provides the functionality of features such as access control, authentication, system management, monitoring, auditing and workflow management. This is achieved via Purview. It allows access control and authentication features such as Azure AD.

ANALYTICS COMPONENT: This section allows the user to run user queries and may also generate structured data for more efficient analysis. It is more suited towards internal users of an enterprise who are able to perform some programming to create custom reports via queries. It is conducted via Azure Synapse Analytics.

Student Name: Mohammed Sadiq Abuwala
Student ID: 45921407

VISUALIZATION COMPONENT: This tier provides the functionalities of visualization via techniques such as line graphs, bar charts, pie charts, scatter plots, etc. The implementation of the visualization component is done via Power BI. It is more suited towards external users and allows data visualization with simple and user friendly techniques.

References

Docs.aws.amazon.com. 2021. *What Is Amazon SageMaker? - Amazon SageMaker*. [online] Available at: <https://docs.aws.amazon.com/sagemaker/latest/dg/whatis.html> [Accessed 23 August 2021].

Docs.microsoft.com. 2021. *Azure AD Connect: Seamless Single Sign-On*. [online] Available at: <https://docs.microsoft.com/en-us/azure/active-directory/hybrid/how-to-connect-sso> [Accessed 23 August 2021].

Elastic. 2021. *What is Elasticsearch?*. [online] Available at: <https://www.elastic.co/what-is/elasticsearch> [Accessed 23 August 2021].

Logz.io. 2021. *Solr vs. Elasticsearch: Who's The Leading Open Source Search Engine?*. [online] Available at: <https://logz.io/blog/solr-vs-elasticsearch/> [Accessed 23 August 2021].

Medium. 2021. *Azure Machine Learning Service - Part 1: An Introduction*. [online] Available at: <https://towardsdatascience.com/azure-machine-learning-service-part-1-an-introduction-739620d1127b#:~:text=Azure%20Machine%20Learning%20(Azure%20ML,and%20model%20development%20skills%20%26%20frameworks.> [Accessed 23 August 2021].

MongoDB. 2021. *Comparing DynamoDB and MongoDB*. [online] Available at: <https://www.mongodb.com/compare/mongodb-dynamodb> [Accessed 23 August 2021].

Student Name: Mohammed Sadiq Abuwala
Student ID: 45921407

MongoDB. 2021. *What Is MongoDB?*. [online] Available at: <https://www.mongodb.com/what-is-mongodb> [Accessed 23 August 2021].

Mysql.com. 2021. *MySQL*. [online] Available at: <https://www.mysql.com/> [Accessed 23 August 2021].

Okta.com. 2021. *Single Sign-On*. [online] Available at: <https://www.okta.com/products/single-sign-on/> [Accessed 23 August 2021].

OneLogin. 2021. *OneLogin vs. Okta: The Difference*. [online] Available at: <https://www.onelogin.com/lp/okta-vs-onelogin> [Accessed 23 August 2021].

Powerbi.microsoft.com. 2021. *What is Power BI | Microsoft Power BI*. [online] Available at: <https://powerbi.microsoft.com/en-us/what-is-power-bi/> [Accessed 23 August 2021].

Support.sas.com. 2021. *SAS Visual Text Analytics*. [online] Available at: <https://support.sas.com/en/software/visual-text-analytics-support.html#:~:text=SAS%20Visual%20Text%20Analytics%20in,of%20your%20text%2Dbased%20data.> [Accessed 23 August 2021].