Storage Solutions for Big Data CA1

Assessment Task

In this CA, you are required to identify a large dataset and perform an analysis using a big data framework for the storage and use an appropriate programming language for the processing.

CA1 Specification

The following elements should be part of the CA1 :

A. Obtain a large dataset from any online public repositories (For example, data.gov.ie, Kaggle and the datasets available at https://hadoopilluminated.com/hadoop_illuminated/Public_Bigdata_Sets.html), while the size of the dataset should justify the complexity level (Structured or Semi-structured or Unstructured). Source data sets can be static (file or database) or use an API to retrieve the data.
B. Utilisation of a distributed data processing environment for some part of the analysis (e.g., Apache Hadoop Map-reduce/ Apache Spark) and storage of data on the hadoop distributed platform.
C. Use any NoSQL database for the storage and processing of queries based on the chosen dataset and objectives of the CA1.
D. Programmatically accessing the source data from the chosen NoSQL database using relevant MapReduce / Spark code.

Scenario

- Initially, the data can be stored into Hadoop/ MySQL/ NoSQL storage. Hadoop MapReduce/ Spark processing would utilise MySQL/ NoSQL database as an input. After processing the data through Mapreduce/ Spark, you can store it into a Hadoop or NoSQL database.
- You can use Python/NumPy/Pandas/Matplotlib to conduct further analysis of the MapReduce output data (e.g., data analysis), and generate data visualisation plots to explain the outcomes.

Deliverables

The results of the data processing activities must be presented in the form of the report and programming code along with comments. This report should highlight and explain the programming and data handling challenges that you will face and the methods you employed to overcome these challenges.

The report should include the following ;

1. A description of the source dataset(s) and the dataset should not be older than five years.
2. A description of the objective of Big data storage and analysis.
3. Details of the data processing activities carried out, including data preparation and processing in the Hadoop MapReduce/ Spark environment.
4. A discussion of the rationale and justification for the choices you have made in terms of data processing, programming language choice, and design patterns that you have implemented.
5. A report of results by making appropriate use of the figures along with captions and tables, etc. Note that MapReduce-style processing in this instance is considered to include platforms, such as Hadoop/ Apache Spark.

Allocation of Marks

The marks for CA1 are based on the following deliverables:

- A clear motivation, objectives and handling of data complexity based on the considered dataset (20%).
- A precise justification of chosen distributed data processing framework as well as deployment (Hadoop or Apache Spark), coding and implementation (30%).
- The results are discussed by supporting appropriate discussion as well as visualisations and conclusions. Harvard style of citations and references should be used in the report (30%).
- The framework and data storage and processing will be demonstrated using a voice-over PowerPoint presentation that lasts no more than seven to eight minutes (20%).

Requirements Summarised

Application Requirements

Input Data:

- Big Data
- static or API

Processing:

- Utilisation of a distributed data processing environment
    - Apache Hadoop Map-reduce/ Apache Spark
- Use relevant MapReduce / Spark code

Storage:

- NoSQL database for the storage and processing of queries

Report Requirements

- Description of the source dataset (no marks allocated).
- Description of the objective of Big data storage and analysis (no marks allocated).
- Motivation, Objectives, and Handling of Data Complexity (20%) .
- Justification of chosen distributed data processing framework, deployment, coding and implementation (30%).
- Discussion as well as visualisations and conclusions. Harvard style of citations (30%).

PowerPoint Presentation

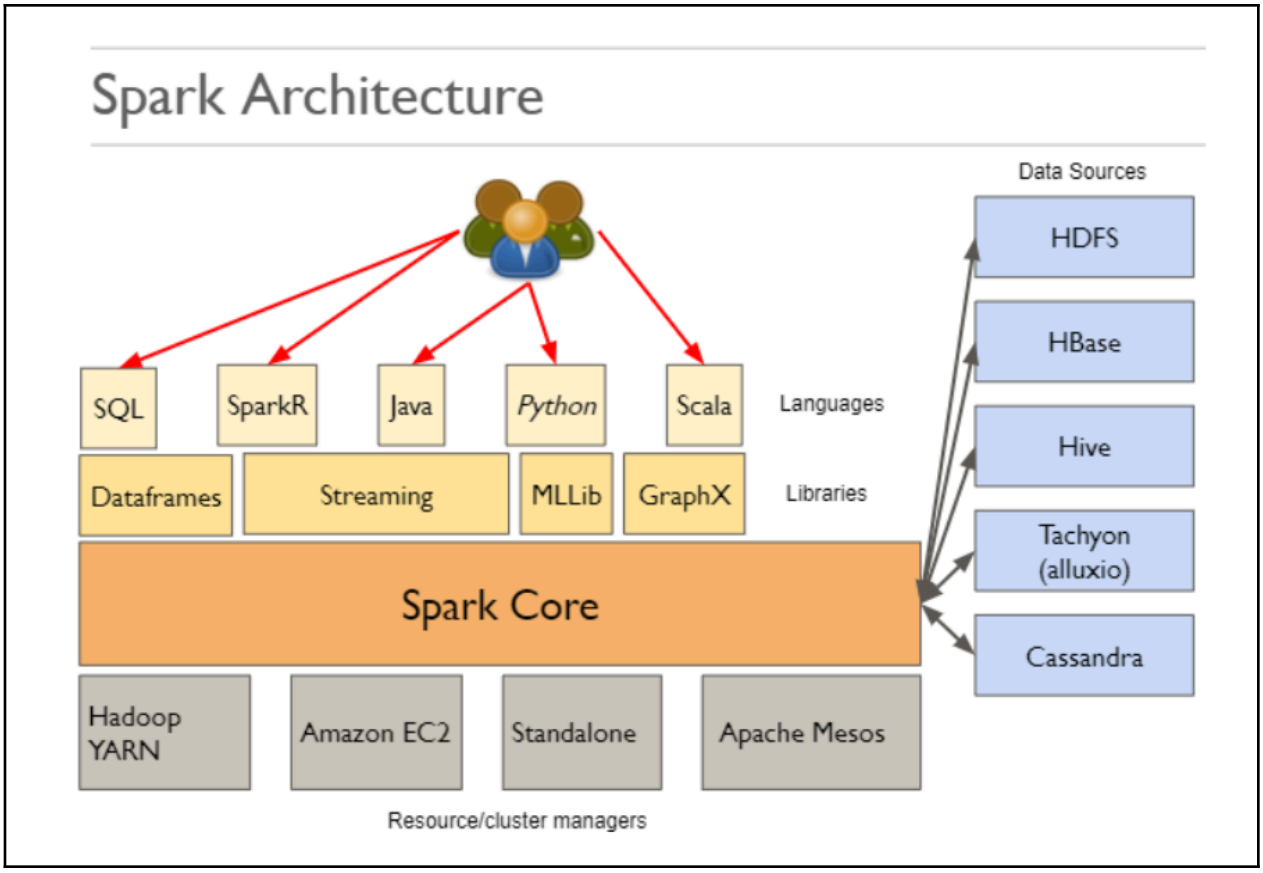
- Framework and data storage and processing.
- Voice-over presentation.
- Seven to eight minutes.

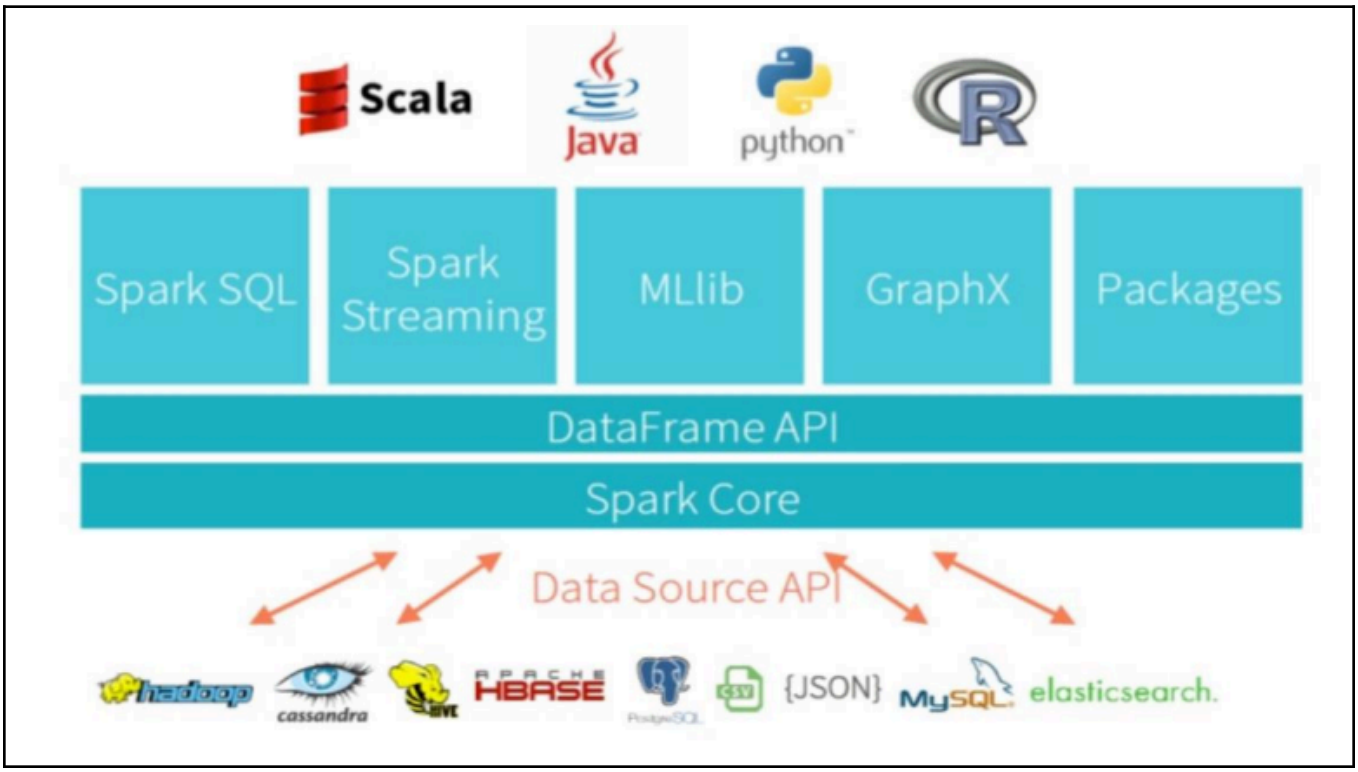
Research and Brainstorming

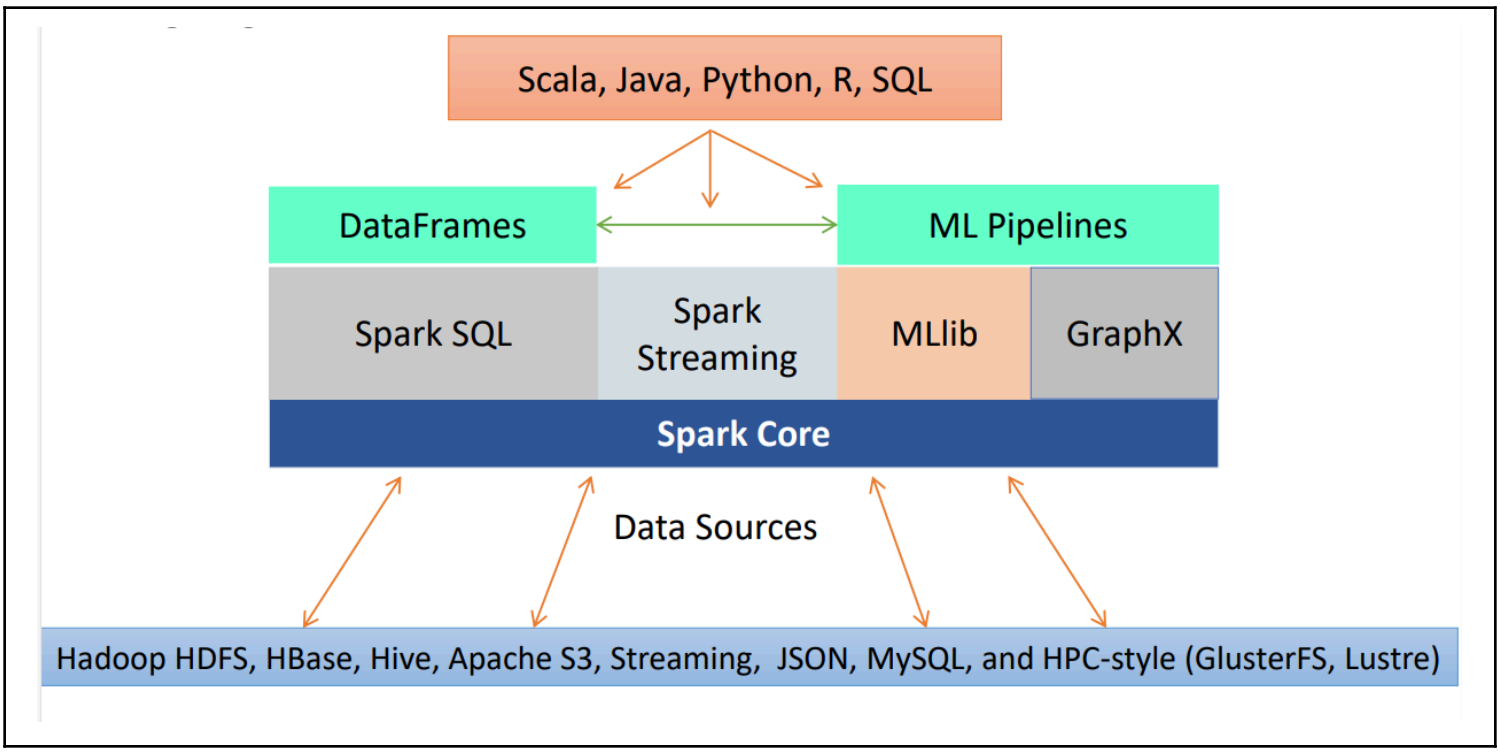
Big Data Applications

1. Real-Time Stream Processing with Kafka and Spark:
    - Set up a real-time data pipeline using Apache Kafka for data ingestion and Spark Streaming for processing.
    - Process live sensor data, stock market updates, or social media streams.
2. Recommendation System using Collaborative Filtering:
    - Build a recommendation engine that suggests relevant products, movies, or music based on user preferences.

    - Implement collaborative filtering algorithms (e.g., user-based or item-based) using tools like Apache Spark.
3. Web Log Analysis with Hadoop and Hive:
    - Analyse web server logs to extract insights.
    - Use Hadoop to process log files and extract relevant information.
    - Create a Hive table to query and analyse the data.
    - Store aggregated results in a database.
4. Predictive Maintenance using Spark MLlib:
    - Predict equipment failures using sensor data.
    - Use Spark MLlib for machine learning.
    - Train models on historical data to predict maintenance needs.
    - Store predictions and maintenance schedules in a NoSQL database like Cassandra.
5. Sentiment Analysis on Social Media Data:
    - Collect a large dataset of social media posts (e.g., tweets) related to a specific topic (e.g., product reviews, political events).
    - Apply natural language processing (NLP) techniques to analyse sentiment, identify trends, and visualise the results.
6. Big Data ETL Pipeline with Spark and MongoDB:
    - Create an ETL (Extract, Transform, Load) pipeline.
    - Extract data from various sources (e.g., CSV files, APIs).
    - Transform and clean the data using Spark.
    - Load the processed data into a MongoDB database.
7. Clickstream Analysis with Hadoop and Impala:
    - Analyse user clickstream data from a website.
    - Use Hadoop to process raw logs.
    - Create an Impala table for querying and analysing clickstream patterns.
    - Store aggregated metrics in a MySQL database.
8. Real-Time Fraud Detection using Spark Streaming and Redis:
    - Detect fraudulent transactions in real time.
    - Stream transaction data using Spark Streaming.
    - Use machine learning models to identify anomalies.
    - Store suspicious transactions in a Redis database.
9. Web Server Log Processing using Hadoop:
    - Analyse web server logs (such as Apache access logs) using Hadoop.
    - Extract relevant information, perform data transformations, and gain insights into user behaviour, traffic patterns, and potential security threats.
10. Predictive Maintenance for Industrial Equipment:
    - Use historical sensor data from industrial machinery (e.g., turbines, pumps) to predict maintenance needs.
    - Apply machine learning models (e.g., regression, time series forecasting) to identify potential failures before they occur.
11. Healthcare Analytics with Electronic Health Records (EHR):
    - Analyze anonymized EHR data to identify patterns related to diseases, treatments, and patient outcomes.
    - Explore correlations, build predictive models, and visualise health trends.


Spark Architecture
Data Sources
HDFS
HBase
Hive
Tachyon (alluxio)
Cassandra
SQL
SparkR
Java
Python
Scala
Languages
Dataframes
Streaming
MLLib
GraphX
Libraries
Spark Core
Hadoop YARN
Amazon EC2
Standalone
Apache Mesos
Resource/cluster managers


Scala
Java
python
R
Spark SQL
Spark Streaming
MLlib
GraphX
Packages
DataFrame API
Spark Core
Data Source API
hadoop
cassandra
HBASE
{JSON}
MySQL
elasticsearch.



Data Sources

1. Kaggle: https://www.kaggle.com/
2. GitHub: https://github.com/
3. Hadoop: Publicly Available Big Data Sets :: Hadoop Illuminated
4. Database Star: https://www.databasestar.com/free-data-sets/
5. DataGov: https://data.gov.ie/

Social Media Analysis

Business Understanding

Sentiment analysis or opinion mining is a type of machine analysis done on a piece of text to determine its sentiments, that is whether it is positive, negative, neutral, or if any other insights could be derived.

Sentiment analysis provides businesses with real-time feedback on marketing campaigns, product launches, and customer experiences. By monitoring sentiment trends, businesses can quickly assess the effectiveness of their strategies and make adjustments as needed to address concerns or capitalise on positive feedback.

1. Get insights about your services or products
Comments on social media, online reviews, blog posts, ratings, feedback, or website content influence customers' purchasing behaviour. Companies can track which products or services are most liked and disliked by analysing the written content based on their emotive language. This can help develop new strategies, especially for the products/services most disliked or least preferred products, which can increase sales revenue.

2. Understand the needs of customers
Understanding whether your customers are satisfied with your services or products is crucial. When customers have more than one bad experience, 80% of them switch to a competitor. So, if the customers have an unpleasant experience, you can detect that quickly through the results of sentiment analysis and meet their expectations and needs.

3. Keep an eye on the competitors
Through sentiment analysis on social media, you can examine the sentiment expressed in the comments, reviews, or feedback that mention competitors to better understand their market share. This can also give you an idea about your company's position in the market. The more information you have about the market, the more action you can take to meet your customers' needs.

4. Develop effective marketing strategies/campaigns
If you are planning a new marketing campaign, then analysing the sentiment related to that topic on social media would give you insights and new ideas for developing effective strategies. For instance, if you are an airline company and want to build a new marketing strategy for international flights, you can analyse reviews' sentiments and understand what would make customers more satisfied.

Methodology

- Ideally we will use unstructured data as the input, that is live streamed using Apache Spark Streaming or Kafka.
- Alternatively we will download a large parquet or csv file that meets the requirements.
- We will initiate analysis using Apache Spark and Anaconda.
- We will perform preprocessing, including some EDA, data preparation and cleaning.
- We will then store this to HDFS.
- We will create a table in Hbase.
- We transfer our data to it from HDFS for storage and further processing.
- We will conduct further cleaning and analysis using PySpark and SparkSQL API's in a Jupyter notebook.
- We will apply word counts, top ten, and sentiment analysis amongst others..
- We will create visuals of analysis and results for the presentation using Pandas, Matplotlib and Seaborn libraries.
- We will document the process and code using screenshots.
- We will report and present our findings.