

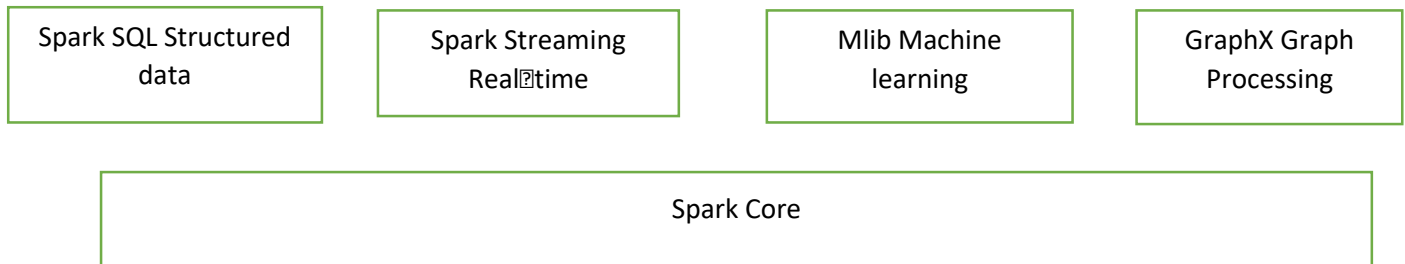
ASSIGNMENT 7

NAME: Aishee Bhattacharya

BATCH: DXC-262-Analytics-B12-Azure

DATE: 07/06/2022

1. Explain what are various components of SPARK with block diagram? explain functionality of every components?



- Spark Core: Spark core is the base engine for large-scale parallel and distributed data processing.
 - Spark SQL: Spark SQL framework component is used for structured and semi-structured data processing
 - Spark Streaming: Spark streaming is a light weight API that allows developers to perform batch processing and real time streaming of data with ease.
 - MLib: M-lib is a low machine library that is simple to use is scalable and compatible with various programming languages.
 - GraphX: GraphX is a spark's own graph computation engine and data store.
2. Explain Spark core in details & how RDD is related to Spark core - explain with Spark program ?

Spark core is the base engine for large-scale parallel and distributed data processing.

It is responsible for

- Memory management
- Fault recovery
- Scheduling, distributing and monitoring jobs on a cluster
- Interacting with storage systems

Spark core is embedded with RDDs an immutable fault tolerant distributed collection of objects that can be operated on in parallel.

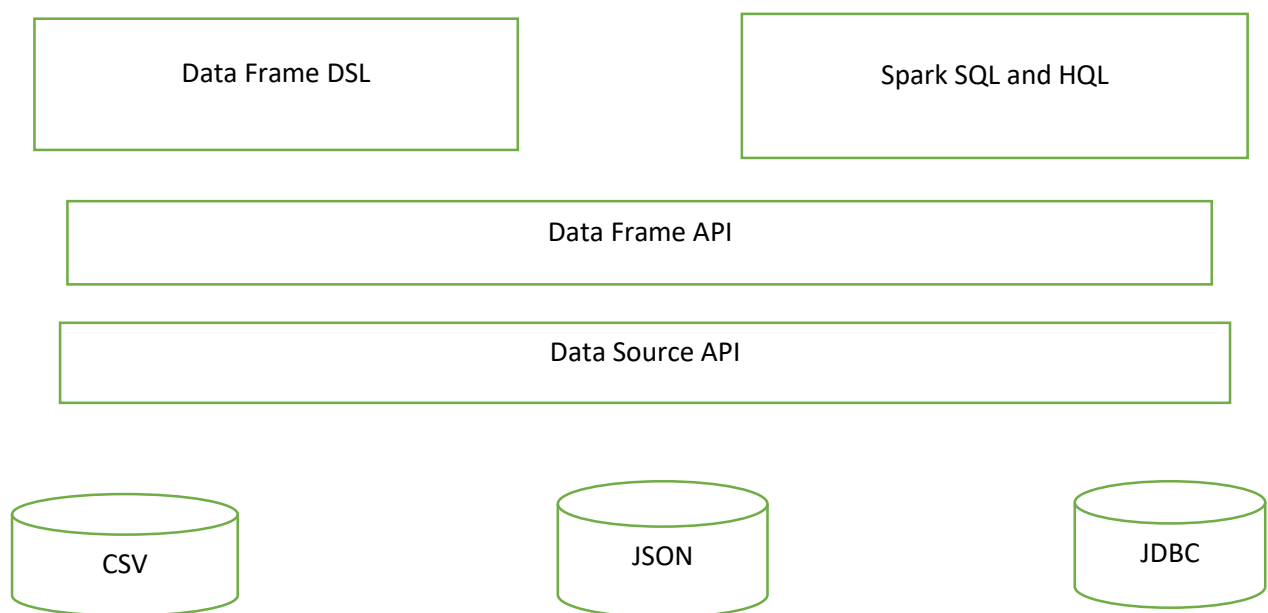
3. Explain various Mlib algorithms Spark is supporting ?

Mlib contains machine learning libraries that have an implementation of various machine learning algorithms.

- Clustering
- Classification
- Collaborative filtering

4. Explain benefits Spark SQL & how relational data will be inserted into SPARK ?

Spark SQL framework component is used for structured and semi-structured data processing.

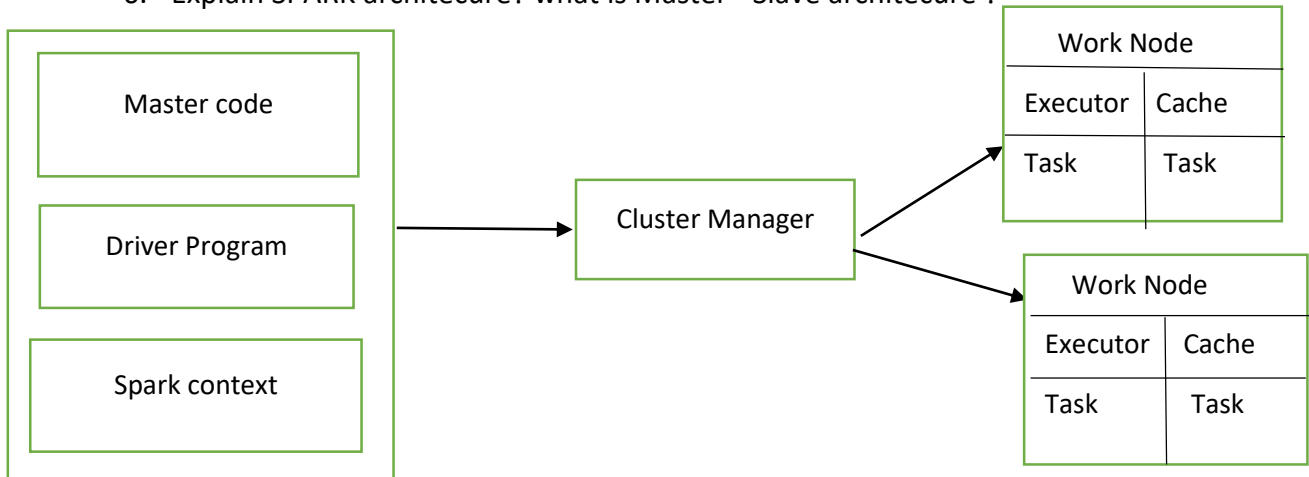


5. Explain Spark streaming in detail ?

Spark streaming is a light weight API that allows developers to perform batch processing and real time streaming of data with ease. It provides secure, reliable and fast processing of live data streams.

Input data stream → Batches of input data → Batches of processed data

6. Explain SPARK architecture? what is Master - Slave architecture ?



7. Explain various cluster managers in SPARK?

- Spark Standalone mode- By default applications submitted to the standalone mode cluster will run in FIFO order, and each application will try to use all available nodes.
- Apache Mesos- It is an open source project to manage computer clusters and also can run Hadoop applications.
- Hadoop YARN- It is the cluster resource manager of Hadoop 2 Spark can be run on YARN.
- Kubernetes-Kubernetes is an open-source system for automating deployment,scaling and management of containerized applications.

8. Explain with sceenshots & steps how to create Cosmos DB ?

Create Azure Cosmos DB Account - Core (SQL)

Basics | Global Distribution | Networking | Backup Policy | Encryption | Tags | Review + create

Azure Cosmos DB is a fully managed NoSQL database service for building scalable, high performance applications. [Try it for free](#), for 30 days with unlimited renewals. Go to production starting at \$24/month per database.

Project Details
Select the subscription to manage deployed resources and costs. Use resource groups like folders to organize and manage all your resources.

Subscription *

Resource Group * [Create new](#)

Instance Details

Account Name *

Location *

Capacity mode ☐ Provisioned throughput ☒ Serverless [Learn more about capacity mode](#)

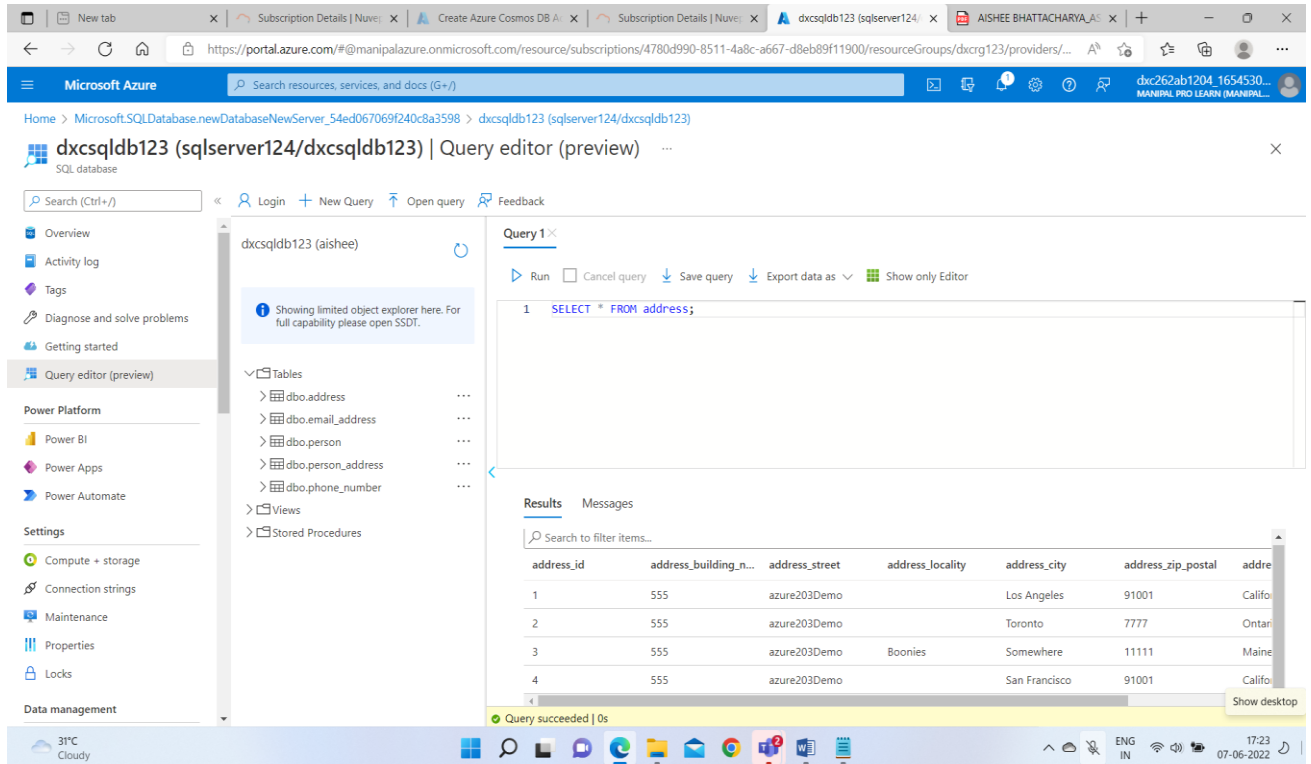
9. Explain with screenshots & step how to insert data into Cosmos DB?

The screenshot shows the Microsoft Azure portal interface for the 'dxccosmosdb2317' account. The 'Data Explorer' view is active, displaying the 'Items' tab for the 'cricketplayers' collection. The 'id' field is highlighted, and the 'Load more' button is visible. The 'New Item' button is also present in the top right.

The screenshot shows the Microsoft Azure portal interface for the 'dxccosmosdb2317' account. The 'Data Explorer' view is active, displaying the 'Items' tab for the 'cricketplayers' collection. The 'id' field is highlighted, and the 'Load more' button is visible. The 'New Item' button is also present in the top right. A JSON document is displayed on the right side of the screen.

```
1 {
2   "01": "Rohit Sharma",
3   "02": "Sachin Tendulkar",
4   "03": "Virat Kohli",
5   "04": "Sourav Ganguly",
6   "05": "MS Dhoni",
7   "06": "Md Shami",
8   "id": "3838f1fc-ee0b-4158-849c-880c64448d17",
9   "_rid": "f-EFA3Fh30MBAAAAAAAAA==",
10  "_self": "db/f-EFA3Fh30MBAAAAAAAAA==/colls/f-EFA3Fh30MBAAAAAAAAA==/",
11  "_etag": "\"01801bca-0000-0100-0000-629ef4d10000\"",
12  "_attachments": "attachments/",
13  "_ts": 1654386833
14 }
```

10. Explain with screenshots & step how to create Azure SQL Db & also explain how to insert data into Azure SQL D?



The screenshot shows the Microsoft Azure portal interface. The top navigation bar includes the Microsoft Azure logo and a search bar. The main content area displays the 'Query editor (preview)' for a SQL database named 'dxcsqldb123 (sqlserver124/dxcsqldb123)'. The left sidebar contains a navigation menu with options like Overview, Activity log, Tags, Diagnose and solve problems, Getting started, Query editor (preview), Power Platform, Power BI, Power Apps, Power Automate, Settings, Compute + storage, Connection strings, Maintenance, Properties, Locks, and Data management.

The 'Query editor (preview)' section shows a query titled 'Query 1' with the following SQL statement:

```
1 SELECT * FROM address;
```

The query has been executed successfully, and the results are displayed in a table. The table has 7 columns: address_id, address_building_n..., address_street, address_locality, address_city, address_zip_postal, and address. The results show 4 rows of data:

address_id	address_building_n...	address_street	address_locality	address_city	address_zip_postal	address
1	555	azure203Demo		Los Angeles	91001	Califo
2	555	azure203Demo		Toronto	7777	Ontar
3	555	azure203Demo	Boonies	Somewhere	11111	Maine
4	555	azure203Demo		San Francisco	91001	Califo

The bottom status bar indicates 'Query succeeded | 0s'.