

Spark Assignment 2

By Himani Anil Deshpande

Spark:

It is an open-source data processing engine which is used for solving large-scale data processing and analytical tasks. **It is a cluster of computing frameworks rather than a database.**

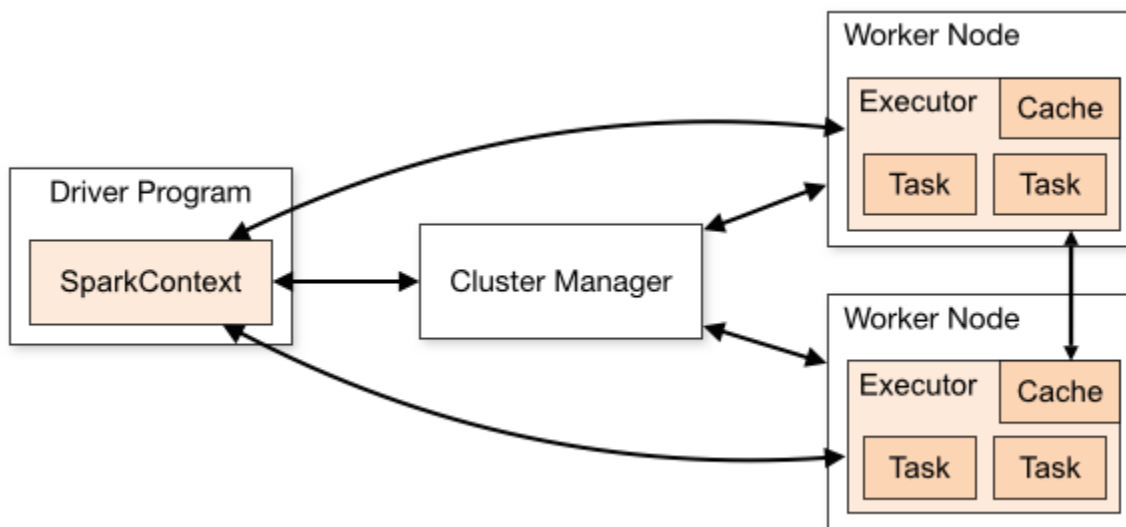
It is a middleware between the data source and the querying frontend.

Its main feature is the in-memory cluster computing that provides incredible processing speed for an application. It contains basic components like fault tolerance, memory management, task scheduling, communicating with data source and its storage systems, etc.

Spark Ecosystem consists of Spark SQL, Spark Streaming, Graph X and core API components for Java, Scala, Python, etc.

- Spark SQL is a domain-specific language for interacting with DataFrames.
- Spark Streaming, which employs mini-batches of data for RDD transformations, allows the same application code that is written for batch analytics to be utilized for streaming analytics as well.
- Spark MLlib is a machine-learning library that simplifies the creation of large-scale machine learning pipelines.
- GraphX, which is an Apache Spark-based distributed graph processing framework.

Architecture:



Spark Assignment 2

By Himani Anil Deshpande

Spark has two main abstractions: Resilient Distributed datasets(RDDs) and Directed Acyclic Graph(DAG)

High level components of Spark Architecture.

Spark Driver:

It maintains the state of the application running on the clusters and also interacts with the cluster manager to get the resources and launch tasks

Spark executors:

They perform the tasks that the driver process commands to do and report back the results to the driver process.

Cluster Manager

There is a built-in Standalone manager that spark has or we can use others like Apache Mesos, Hadoop YARN, kubernetes, etc.

The cluster manager takes care of all the machines in the cluster and executes the spark applications.

It has its own worker and master abstractions/nodes. The cluster manager is responsible for efficiently scaling up the compute nodes

Spark can load data from a file system or database and create a RDD to store and process it. The RDDs are immutable and are useful for parallel computation and transformations of data. They are highly resilient so that they can recover if anything fails. The data in RDD is split into blocks using a key.

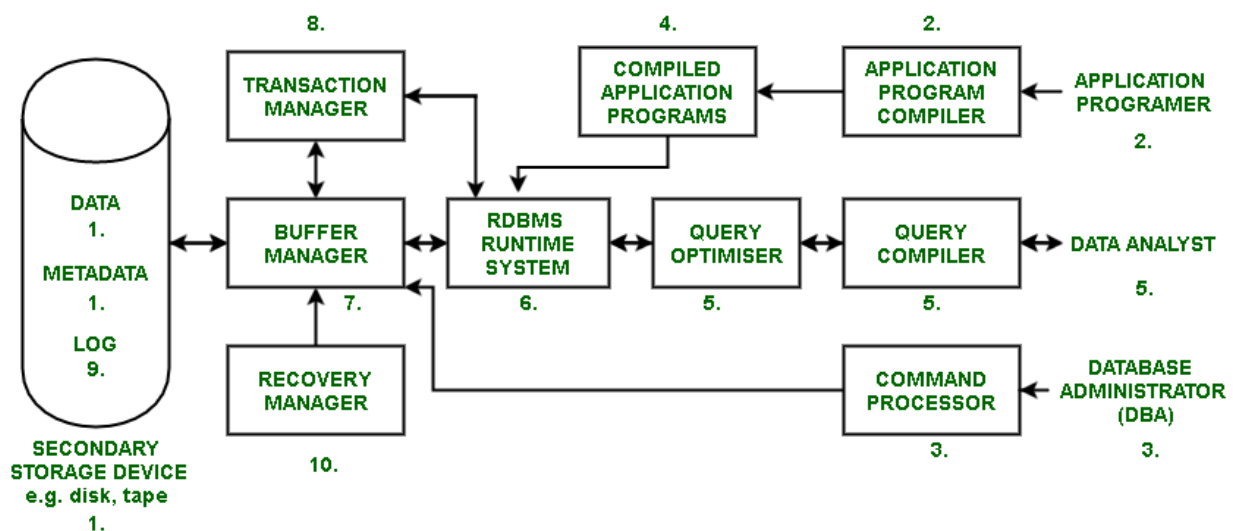
Spark Assignment 2

By Himani Anil Deshpande

RDBMS

It is a type of **database management system** which has functions and capability to maintain security, integrity, accuracy and consistency of data. It stores the data in a structured format using tables using the most basic functions of CRUD(create, read, update and delete), which are Data manipulation techniques/language

Architecture:



1. Uses SSD(Secondary Storage Devices) like Diska, tapes, etc for storing data, metadata and logs. The Application programs written in Java, C/C++, Python, etc are used for communicating with databases using SQL
2. A compiler in an RDBMS transforms SQL commands into lower-level language, executes them, and saves them to a secondary storage medium.
3. The DBA sets up the database and uses DDL(Data Definition Language) for creating, dropping tables, access controls, constraints, relationships through foreign keys, and adding columns, etc using command processors.
4. Using a compiler the application programs are compiled to create executable files and store them on SSDs.
5. Query compiler and Optimizer are used for manipulating the data in DB.
6. RDBMS Run time system interacts with transaction and buffer manager for running the compiled queries and application programs.
7. The data is temporarily stored in main memory by buffer manager and for better memory management and faster performance paging algorithms are used.
8. Transactional Manager looks after the Atomicity property, which is doing a transaction completely or not at all.

Spark Assignment 2

By Himani Anil Deshpande

9. Log is for keeping track of all transactions so that when a system fails the partial transactions can be finished or recalled.
10. Recovery manager works in tandem with log system so that whenever the system fails it can be taken back or restored to the steady state.

Other points on Data Access and other properties

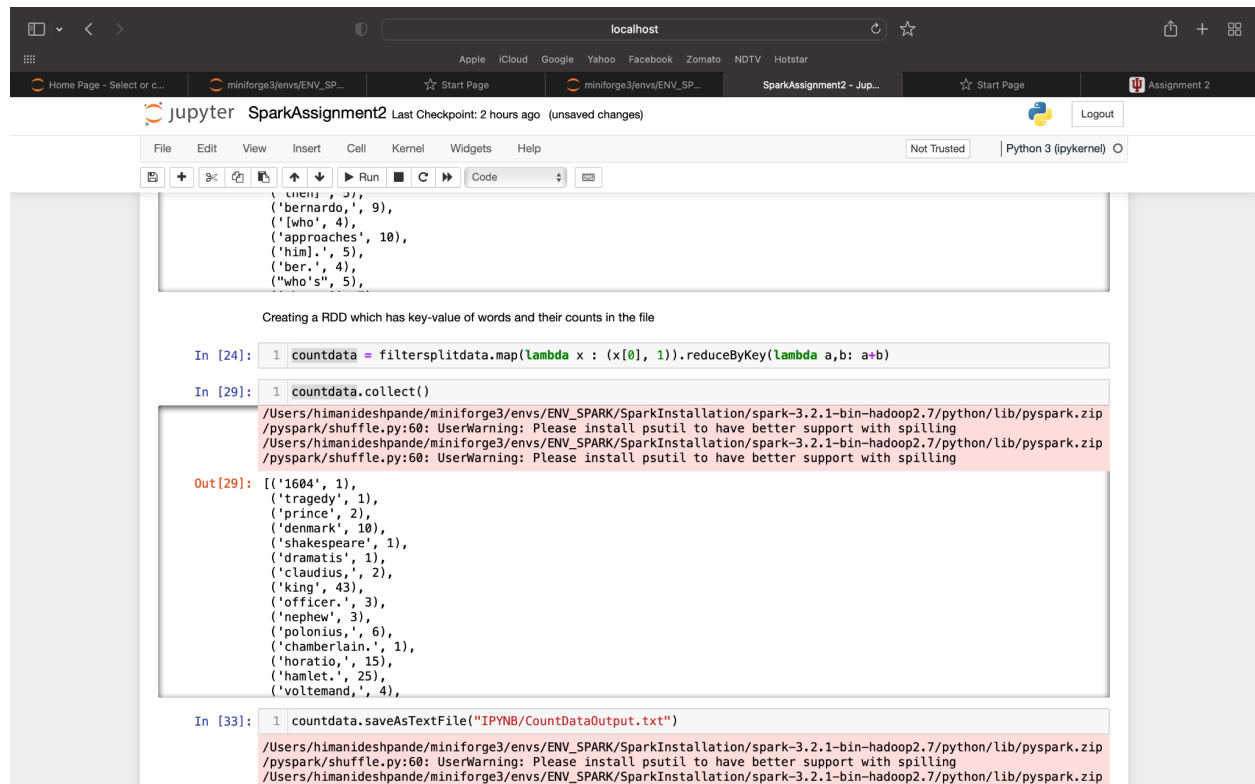
RDBMS work better with **low data load** i.e Gigabytes of Data. However, it is not competent enough to handle data loads in **Tera and Petabytes**., which Spark can handle.

RDBMS can handle only **structured and semi-structured** data. Spark can handle **unstructured** data apart from the semi and structured data.

Spark offers **real time computation and low latency** and it has a programming layer which gives powerful caching and **disk persistence** capabilities.

RDBMS has good latency but it does not give a **quick response as Spark** if the data is huge. So Spark has higher latency and throughput compared to RDBMS.

Output:



The screenshot shows a Jupyter Notebook titled "SparkAssignment2" running on a local host. The notebook contains the following code and output:

```
def countdata = filtersplitdata.map(lambda x : (x[0], 1)).reduceByKey(lambda a,b: a+b)
```

```
In [24]: 1 countdata = filtersplitdata.map(lambda x : (x[0], 1)).reduceByKey(lambda a,b: a+b)
```

```
In [29]: 1 countdata.collect()
```

```
Out[29]: [('1604', 1), ('tragedy', 1), ('prince', 2), ('denmark', 10), ('shakespeare', 1), ('dramatis', 1), ('claudius', 2), ('king', 43), ('officer', 3), ('nephew', 3), ('polonius', 6), ('chamberlain', 1), ('horatio', 15), ('hamlet', 25), ('voltmand', 4)]
```

```
In [33]: 1 countdata.saveAsTextFile("IPYNB/CountDataOutput.txt")
```

The output of the `collect()` command is a list of tuples representing word counts. The output of the `saveAsTextFile()` command is a warning message: "UserWarning: Please install psutil to have better support with spilling".

Spark Assignment 2

By Himani Anil Deshpande

References:

<https://www.geeksforgeeks.org/rdbms-architecture/>

<https://medium.com/edureka/spark-architecture-4f06dcf27387>

<https://www.analyticsvidhya.com/blog/2020/11/data-engineering-for-beginners-get-acquainted-with-the-spark-architecture/>