EXPERIMENT - 9A

AIM: Case study on Apache Spark and Power BI

THEORY:

APACHE SPARK

Spark is an Apache project advertised as “lightning fast cluster computing”. It has a thriving open-source community and is the most active Apache project at the moment. Spark provides a faster and more general data processing platform. Spark lets you run programs up to 100x faster in memory, or 10x faster on disk, than Hadoop. Last year, Spark took over Hadoop by completing the 100 TB Daytona GraySort contest 3x faster on one tenth the number of machines and it also became the fastest open source engine for sorting a petabyte. Spark also makes it possible to write code more quickly as you have over 80 high-level operators at your disposal. Another important aspect when learning how to use Apache Spark is the interactive shell (REPL) which it provides out-of-the box. Using REPL, one can test the outcome of each line of code without first needing to code and execute the entire job. The path to working code is thus much shorter and ad-hoc data analysis is made possible.

Spark was introduced by Apache software system Foundation for rushing up the Hadoop procedure computing software system method. As against a standard belief, Spark isn't a changed version of Hadoop and isn't, really, keen about Hadoop because it's its own cluster management. Hadoop is just one of the ways in which to implement Spark. Spark uses Hadoop in 2 ways in which – one is storage and second is process. Since Spark has its own cluster management computation, it uses Hadoop for storage purposes solely. The most feature of Spark is its in memory cluster computing that will increase the processing speed of Associate in Nursing application. Spark is meant to cover a large variety of workloads like batch applications, unvarying algorithms, interactive queries and streaming. except supporting these workloads in an exceedingly various system, it reduces the management burden of maintaining separate tools.

Spark also has the following key features:

- APIs are currently available in Scala, Java, and Python, with support for additional languages (such as R) on the way.
- Compatible with the Hadoop ecosystem and data sources (HDFS, Amazon S3, Hive, HBase, Cassandra, etc.)
- Can run on Hadoop YARN or Apache Mesos clusters, as well as standalone.

Advantages of Apache Spark:

- **Speed**: Spark helps to run applications in Hadoop cluster, up to one hundred times quicker in memory, and ten times quicker once running on disk. This is often doable by reducing the range of read/write operations to disk.This helps to scale back most of the disc browse and write – the most time consuming factors of the information process.

- **Combines SQL, Streaming, and Complex Analytics**: In addition to straightforward "map" and "reduce" operations, Spark supports SQL queries, streaming knowledge, and complicated analytics like machine learning and graph algorithms out-of-the-box. Not solely that, users will mix these capabilities seamlessly during a single workflow.

- **Supports Multiple Languages:** Spark provides constitutional APIs in Java, Scala, or Python. Therefore, you'll be able to write applications in several languages. Spark comes up with eighty high-level operators for interactive querying.

- **Advanced Analytics:** Spark not solely supports 'Map' and 'reduce'. It conjointly supports SQL queries, Streaming knowledge, Machine learning (ML), and Graph algorithms.

- **Runs Everywhere**: Spark runs on Hadoop, Mesos, standalone, or within the cloud. It will access various information sources as well as HDFS, Cassandra, HBase etc.

- **Deep Learning Pipelines**: Apache Spark supports deep learning via Deep Learning Pipelines. mistreatment of the prevailing pipeline structure of MLlib, you'll be able to decide into lower level deep learning libraries and construct classifiers in only many lines of code, further as apply custom TensorFlow graphs or Keras models to incoming knowledge. These graphs and models will even be registered as custom Spark SQL UDFs (user defined functions) so the deep learning models are applied to knowledge as a part of SQL statements.

Components of Spark:

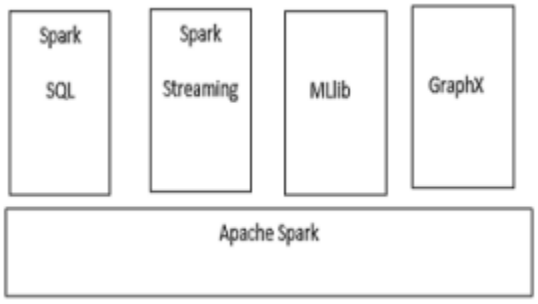
The **Spark Core** engine is the foundation for large-scale parallel and distributed data processing. It is in charge of: Memory administration and fault recovery interacting with storage systems by scheduling, distributing, and monitoring jobs on a cluster. Spark introduces the concept of an RDD (Resilient Distributed Dataset), which is an immutable, fault-tolerant, distributed collection of objects that can be worked on concurrently. An RDD can contain any type of object and is created by loading a dataset from an external source or distributing a collection from the driver programme.

SparkSQL is a Spark component that supports querying data either via SQL or via the Hive Query Language. It originated as the Apache Hive port to run on top of Spark (in place of MapReduce) and is now integrated with the Spark stack. In addition to providing support for various data sources, it makes it possible to weave SQL queries with code transformations which results in a very powerful tool.

MLlib may be a distributed machine learning framework on top of Spark as a result of the distributed memory-based Spark design. It is, in step with benchmarks, done by the MLlib developers against the Alternating method of least squares (ALS) implementations. Spark MLlib is ninefold as quick because the Hadoop disk-based version of Apache driver (before driver gained a Spark interface).

GraphX is a distributed graph-processing framework on high of Spark. It provides Associate in Nursing API for expressing graph computation that may model the user-defined graphs by mistreatment of the Pregel abstraction API. It conjointly provides Associate in Nursing optimized runtime for this abstraction.

Spark Streaming supports real time processing of streaming data, such as production web server log files (e.g. Apache Flume and HDFS/S3), social media like Twitter, and various messaging queues like Kafka. Under the hood, Spark Streaming receives the input data streams and divides the data into batches.



Apache Spark Operations RDD and MapReduce:

Resilient Distributed Datasets (RDD) could be a basic arrangement of Spark. It is an immutable distributed assortment of objects. every dataset in RDD is split into logical partitions, which can be computed on totally different nodes of the cluster. RDDs will contain any variety of Python, Java, or Scala objects, together with user-defined categories. Formally, associate degree RDD could be a read-only, divided collection of records. RDDs may be created through settled operations on either information on stable storage or different RDDs.

Data Sharing is Slow in MapReduce: MapReduce is widely adopted for process and generating massive datasets with a parallel, distributed rule on a cluster. It permits users to write parallel computations, employing a set of high level operators, while not having to stress regarding work distribution and fault tolerance.

Data Sharing using Spark RDD: Data sharing is slow in MapReduce due to replication, publishing, and disk IO. Most of the Hadoop applications, they pay over 90% of the time doing HDFS read-write operations.

Other Apache Spark Use Cases

- ➢ Potential use cases for Spark include detection of earthquakes.
- ➢ In the game industry, processing and discovering patterns from the potential firehose of real-time in-game events and being able to respond to them immediately is a capability that could yield a lucrative business, for purposes such as player retention, targeted advertising, auto-adjustment of complexity level, and so on.
- ➢ In the e-commerce industry, real-time transaction information could be passed to a streaming clustering algorithm like k-means or collaborative filtering like ALS
- ➢ In the finance or security industry, the Spark stack could be applied to a fraud or intrusion detection system or risk-based authentication.
- ➢ It could achieve top-notch results by harvesting huge amounts of archived logs, combining it with external data sources like information about data breaches and compromised accounts and information from the connection/request such as IP geolocation or time.

POWER BI

Microsoft Power BI is a data visualization platform that is primarily used for business intelligence. Power BI's dashboard, which is intended for use by business professionals with varying levels of data knowledge, is capable of reporting and visualizing data in a variety of formats, including graphs, maps, charts, scatter plots, and more.

Power BI is made up of several interconnected applications, including Power BI Desktop, Pro, Premium, Mobile, Embedded, and Report Server. While some of these apps are free to use, paid subscriptions to the pro and premium versions offer enhanced analytics capabilities.

- Pre-built dashboards and reports for SaaS Solutions
- Power BI allows real-time dashboard updates.
- Offers Secure and reliable connection to your data sources in the cloud or on-premises
- Power BI offers Quick deployment, hybrid configuration, and secure environment.
- Allows data exploration using natural language query
- Offers features for dashboard visualization regularly updated with the community.

Power BI Tools:-

Power BI Desktop

Power BI desktop is the primary authoring and publishing tool for Power BI. Developers and power users use it to create brand new models and reports from scratch.
Costs: Free

Power BI service

Online Software as a Service (SaaS) where Powe Bl data models, reports, dashboards are hosted. Administration, sharing, collaboration happens in the cloud.
Pro license: $10/users/month

Power BI Data Gateway

Power BI Data Gateway works as the bridge between the Power Bl Service and on-premise data sources like DirectQuery, Import, Live Query. It is Installed by Bl Admin.

Power BI Report Server

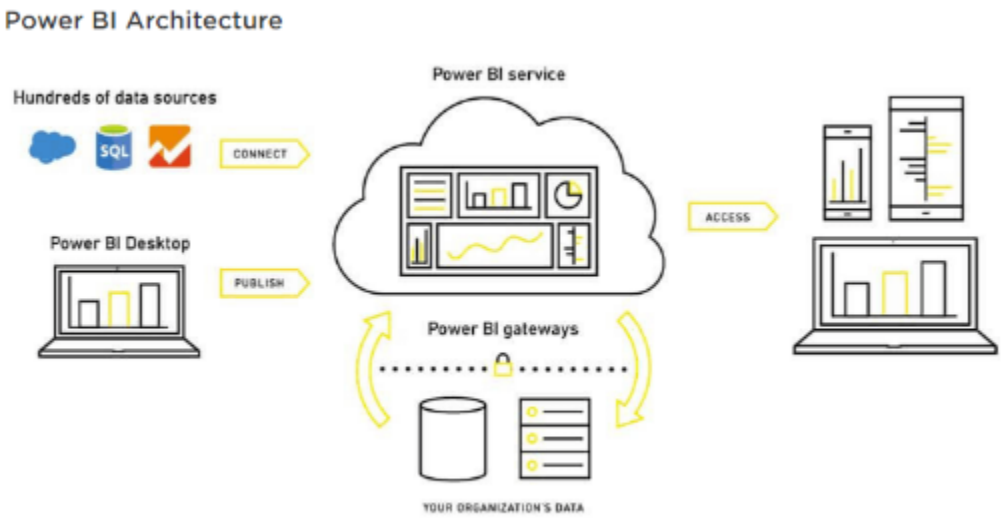
It can host paginated reports, KPIs, mobile reports, & Power Bl Desktop reports. It is updated every 4 months and installed/managed by the IT team. The users can modify Power Bl reports other reports created by the development team.

Power BI Mobile Apps

Power BI mobile app is available for iOS, Android, Windows. It can be managed using Microsoft Intune. You can use this tool to view reports and dashboards on the Power Bl Service Report Server.

Key terms used in Power BI

Architecture of Power BI



Architecture of Power BI

Advantages of Power BI

Here are the advantages of using Power BI:

- Offers pre-built dashboards and reports for SaaS Solutions
- Provide real-time dashboard updates.
- Secure and reliable connection to your data sources in the cloud or on-premises
- Power BI offers quick deployment, hybrid configuration, and a secure environment.
- Data exploration using natural language query.

Disadvantages of Power BI

Here, are Cons/drawbacks of using Power BI:

- Dashboards and reports only shared with users having the same email domains.
- Power Bl will not mix imported data, which is accessed from real-time connections.
- Dashboards never accept or pass user, account, or other entity parameters.
- Very few data sources that permit real-time connections to Power BI reports and dashboard.

CONCLUSION: To summarize, Spark aids in the challenging and computationally intensive task of processing large volumes of real-time or archived structured and unstructured data. Power BI is a data visualization platform that is primarily used for business intelligence. Power BI's dashboard, which is intended for use by business professionals with varying levels of data knowledge, is capable of reporting and visualizing data. Thus we studied about Power BI and Apache Spark.