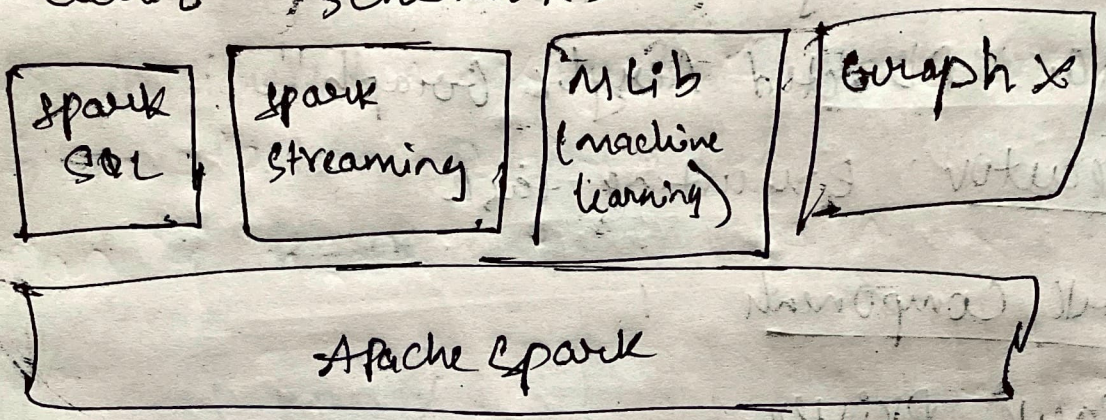
SPARK SQL

→ It is a Spark module for structured data processing

→ It is a component on top of spark core that introduces a new data abstraction called "SchemaRDD"



Challenges:

→ Perform ETL to and from various data sources

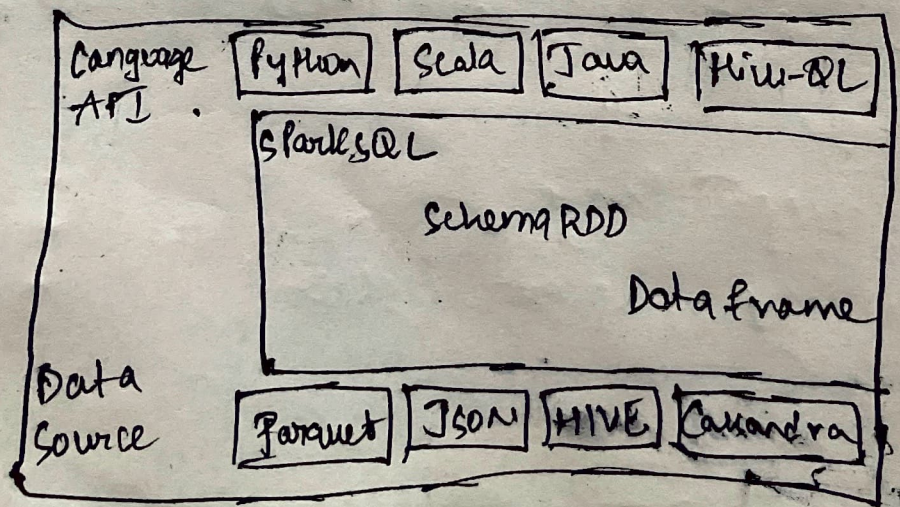
→ Perform Advanced analytics

Solutions

→ A Dataframe API that can perform relational operations.

→ A highly extensible optimizer, catalyst which uses Scala to add rule etc.

Architecture:



Features:

→ Integrated:

① Integrated API's in Python, Scala, Java

② easy to run SQL queries alongwith algorithms.

→ Unified Data Access:

① Load & query data from various Process

② schema RDD provide a single interface for efficiently working with structured data.

→ Hive compatibility

① Run unmodified queries on warehouse

② Simply install it alongside Hive

→ Scalability

① Connect through JDBC & ODBC uses the same engine for both queries

② takes advantage of RDD Model to support Mid-query fault tolerance.

SPARK RDD

→ RDD is a fundamental data structure of spark

→ It is immutable collection of objects that can be stored in memory.

→ RDD is divided into logical partitions

→ Parallel functions transformations

⇒ Automatically rebuilt on failure

Dataset and Data frame :-

⇒ A distributed collection of data :, into named columns.

⇒ Equivalent to relational tables

⇒ The API is designed for modern big data & data science.

Plan optimization & Execution