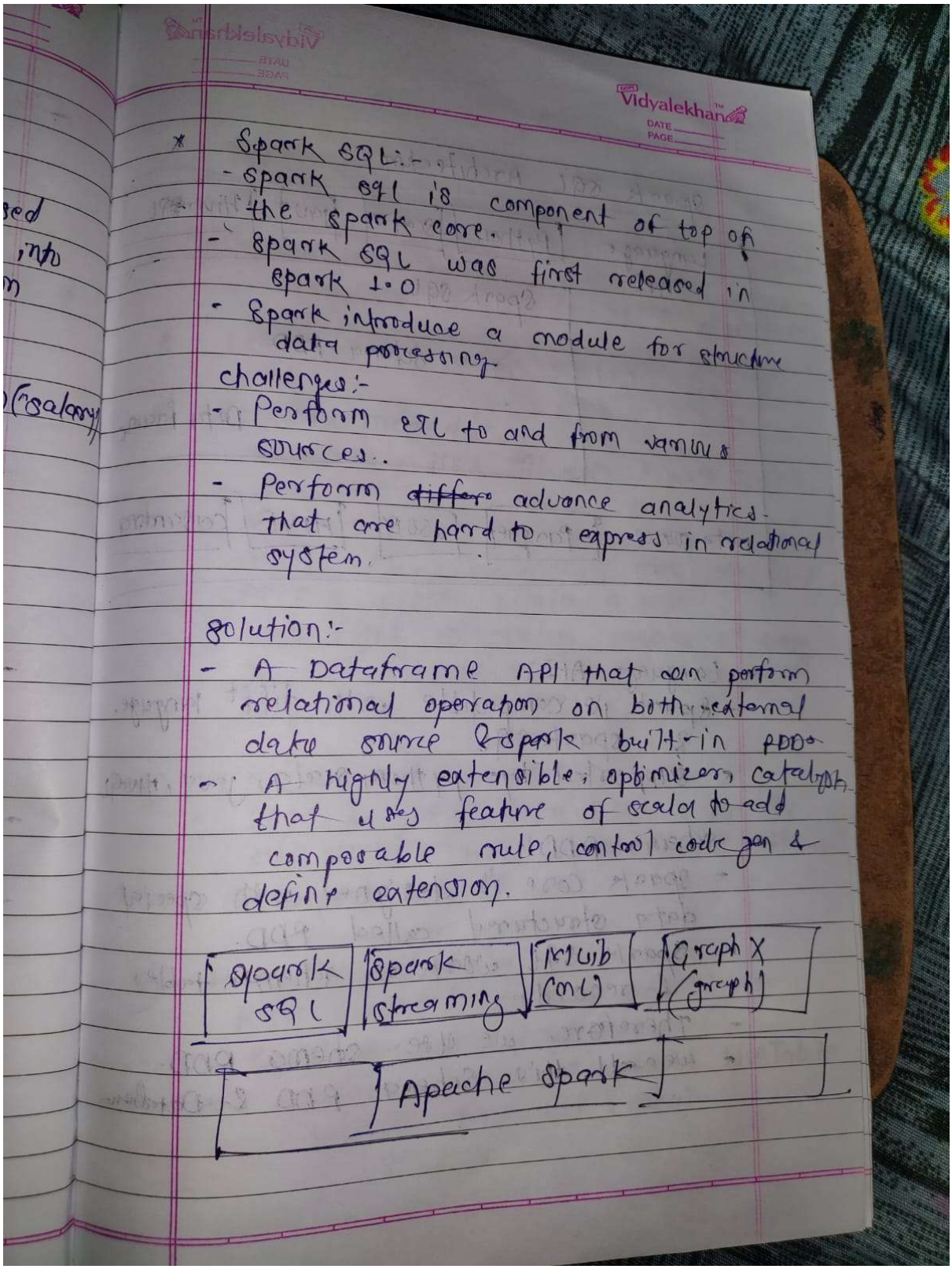
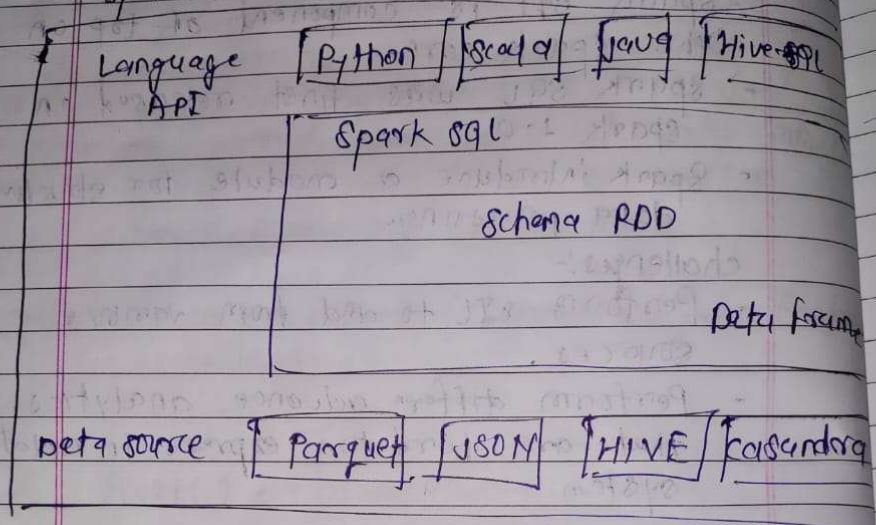
Name: Krushnakumar Patle

Email: krishnapatle128@gmail.com

Batch: Data Engineering Batch-1



Spark SQL Architecture




- Language API :-
  - Spark is compatible with diff language & spark SQL.
  - supported by - python, scala, java, HiveQL
- Schema RDD :-
  - Spark Core is designed with special data structured called RDD.
  - Spark SQL work on schema, table, & records.
  - Therefore, we use schema RDD.
  - we call this schema RDD & Dataframe

1 Data source :-
- It is text file, Avro file etc.
- Data source for spark SQL is diff
- Those are Parquet file, JSON documen, HIVE tables, & casandra database.

4 features of spark SQL:-

1) It Integrated :-
- seamlessly mix SQL queries with spark program.
- spark SQL lets you query struct'd data as a RDD in spark, with integrated API in python, scala, & java.
- This tight integration makes it easy to run SQL queries alongside analytical alg.

2) Unified data access :-
- Load & query data from a variety of source. like Apache Hive table, Parquet.

3) Hive Compatibility :-
- Run unmodified Hive queries on existing warehouse.

ex. SELECT COUNT (*) FROM hiveTable WHERE hive_udf (data)

* Hive:-
Apache hive is a distributed, fault-tolerant data warehouse system that enables analytics at a massive scale.

* User-Defined Functions (UDF):-
- UDF:- Plug in your own processing code & invoke it from Hive query.
- UDF (Plain UDF)
  - Input: single row, output: single row
- UDAF (User-defined Aggregate fun)
  - Input: multiple rows, Output: single row

4) Standard Connectivity:
- Connect through JDBC or ODBC

5) Scalability:-
- Use same engine for both interactive & long queries.
- Do not worry about using a diff engine for historical data.

Vidyalekhan
DATE
PAGE

* Spark RDD (Resellrent Distributed Dataset)
- RDD is fundamental data structure of spark
- It is an immutable distributed collection of objects that can be stored in memory or disk across a cluster
- Each dataset in RDD is divided into logical partitions which may be computed on diff nodes of the cluster
- Automaticall rebuilt on failure
- RDDs can certain any type of python, java, or scala objects.
- Read only partitioned collection of record.
- created through deterministic operation on either data on stable storage or others RDDs.
- RDD is fault-tolerant collection of element that can be operated on in parallel.
- Two ways to create RDD:-
  (1) parallelizing;
  (2) referencing a dataset in an external storage system

* Dataset & Dataframe:-
- A distributed collection of data, which is organized into named columns.
- Dataframe can be constructed from an array of diff sources such as Hive table, Structure Data files.

* Dataframe:-
Data is organised into named column like a table in a relational database

* Dataset: a distributed collection of data

* features of Dataset / Dataframe:-
- Ability to process the data in Kb to Pb
- support diff data format.
- can be easily integrated with all Big data tools & frameworks
- provides API for python, Java, Scala, & R programming

Plan optimization & Execution




Dataframes & SQL share the same optimization/execution pipeline