CS523 - BDT
Big Data Technologies

Final Project

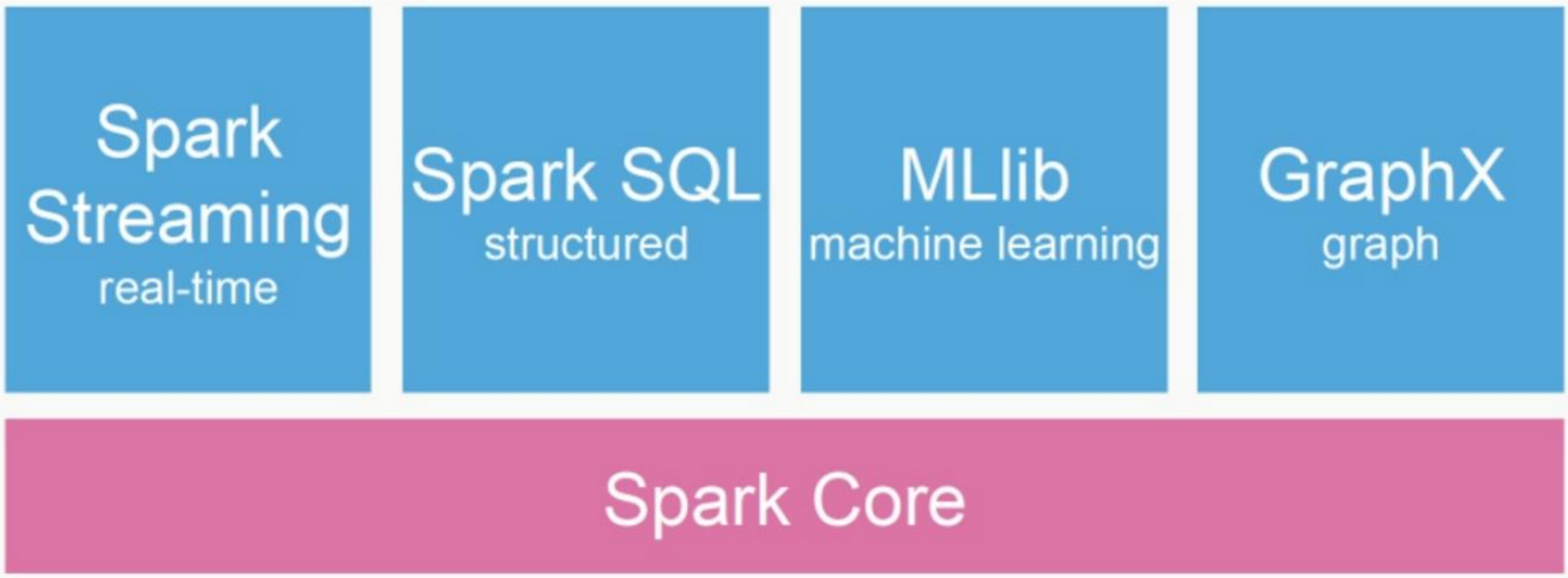
(Knowing and Showing your full potential)

Project Details

Project Parts	Points	Sakai Submission Due Date
1, 2	7 each	20th Mar, Tuesday till 10 pm
3	4	20th Mar, Tuesday till 10 pm
4	4	20th Mar, Tuesday till 10 pm
5	3	19-20 Mar, Monday and Tuesday

- Team size <= 2 students
- Each team will have a short presentation and demo of project parts 1, 2, 3 & 4 (20-25 mins) on Mar 19th & 20th.

Spark Ecosystem Projects



Spark Streaming

- For Real-time predictions and recommendations.
- Spark streaming lets users run their code over a small piece of incoming stream of data in a scale.
- Use cases for Spark Streaming:
  - You just walk by the Walmart store and the Walmart app sends you a push notification with a 20% discount on your favourite clothing brand.
  - Uber – Every day this multinational online taxi dispatch company gathers terabytes of event data from its mobile users. By using Kafka, Spark Streaming, and HDFS to build a continuous ETL pipeline, Uber can convert raw unstructured event data into structured data as it is collected, and then use it for further and more complex analytics.
  - For a stream of weblogs, if you want to get alerts within seconds- Spark Streaming is helpful.

Spark SQL

- Provides functions for manipulating large sets of distributed, structured data using a SQL subset supported by Spark and Hive SQL (HiveQL).
- Used for reading and writing data to and from JSON files, Parquet files, Avro files, RDBMSs, Hive, and others.
- Seamlessly mix SQL queries with Spark programs.
- Operations on DataFrames and DataSets at some point translate to operations on RDDs and execute as ordinary Spark jobs.
- Access records in HBase table with SQL query.
- Run unmodified Hive queries on existing data.
- Connect through JDBC or ODBC using Thrift server.

Data Visualization

- Big Data is made of numbers and numbers are difficult to look at.
- Because of the way the human brain processes information, using charts or graphs to visualize large amounts of complex data is easier than poring over spreadsheets or reports.
- Data visualization is a quick, easy way to convey concepts in a universal manner – and you can experiment with different scenarios by making slight adjustments.
- Data visualization is the presentation of data in a pictorial or graphical format. It enables decision makers to see analytics presented visually, so they can grasp difficult concepts or identify new patterns.
- Data visualization can also:
  - Identify areas that need attention or improvement
  - Clarify which factors influence customer behavior
  - Help you understand which products to place where
  - Predict sales volumes

Project Parts Details

- **Part 1.** [7 points] **Create your own project for Spark Streaming.**
  - ✓ Remember, it should be interesting and useful.
  - ✓ Provide detailed instructions.
- **Part 2.** [7 points] **Create your own project using Spark SQL and Hbase/Hive together.**
  - ✓ Provide detailed instructions.
- **Part 3.** [4 points] **In any of the parts 1 and 2 above, show the proper use of any of the data visualization tools like** Tableau, Jupyter, Plotly, **etc.**
- **Part 4.** [4 points] **Do some research and create a simple demo project for any one of the following tools:** Presto, Impala, Phoenix, Storm, Kafka
- **Part 5.** [3 Points] **In class Presentation of all the 4 parts. Be professional!**
  - ✓ Submit your Presentation in Sakai with the Project.

Public Datasets

- Amazon Web Services
- UCI Machine Learning Repository
- Kaggle
- Data Science Central

What to Submit

- All the .java and .class files
- Bash script files for each project part where in I should be able to find all the commands to run your applications.
- All the input files and output files generated after running the program
- *Readme* file explaining the details of parts 1, 2, 3 & 4.
- Submit a .zip file of all the above mentioned documents.