CS523 - BDT
Big Data Technology

Final Project

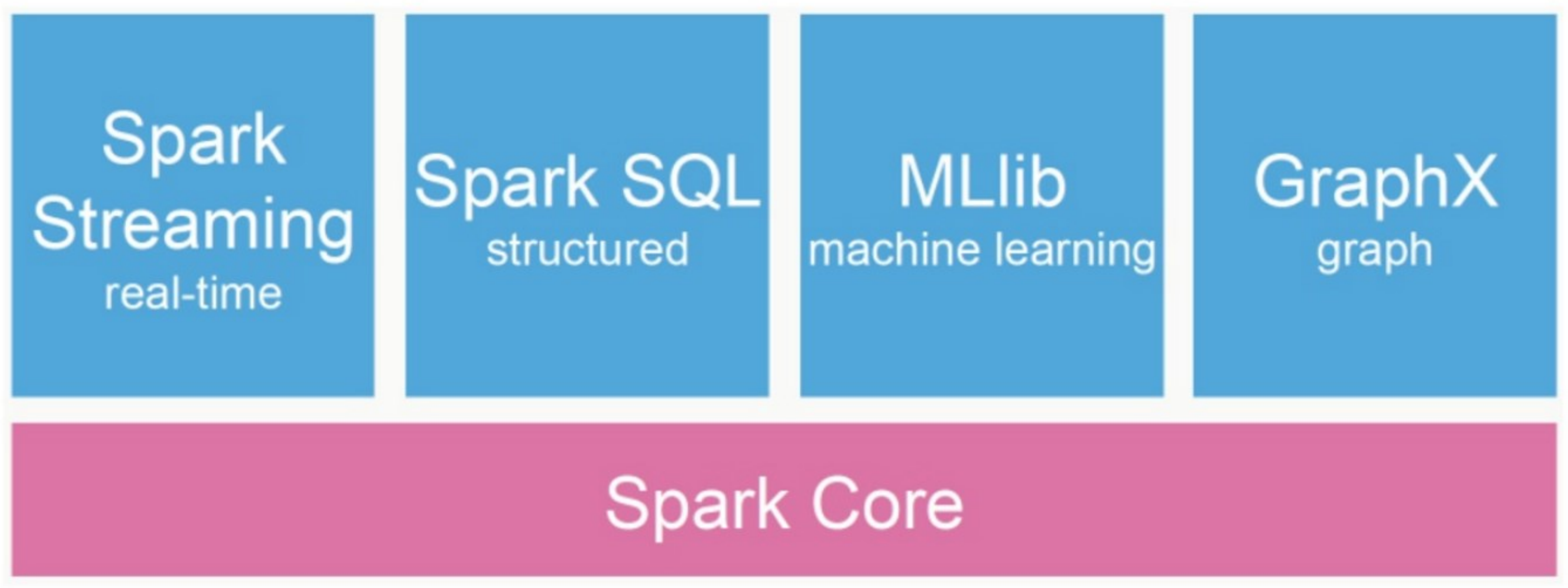
(Knowing and Showing Your Hidden Potential)

Project Details

Project Parts	Points	Sakai Submission Due Date
1, 2	5 each	Apr 25, Tuesday till 10 pm
3	4	
4	4	
5	2	Apr 24 & 25, Monday & Tuesday
Total	**20**	**20% of the total course grade**

- Max 4 students in a Team
- Each team will have a short presentation and demo of project parts 1, 2, 3 & 4 (20-25 mins) on April 24th & 25th.

Spark Ecosystem



Spark Streaming
real-time

Spark SQL
structured

MLlib
machine learning

GraphX
graph

Spark Core

Spark Streaming

- Spark Streaming is a scalable, high-throughput, fault-tolerant stream processing module for live data streams – Used 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.

- Data ingestion can be done from many sources like Kafka, Flume, Amazon Kinesis or TCP sockets and processing can be done using complex algorithms that are expressed with high-level functions like map, reduce, joins, etc.

- Finally, processed data can be pushed out to filesystems, databases and live dashboards.

Spark Streaming contd..

- Data stream is divided into batches called **DStreams**, which internally is a sequence of RDDs. The RDDs are then processed using Spark APIs, and the results are returned in batches.
- Spark Streaming maintains a state based on data coming in a stream and this is called as stateful computations.
- It also allows window operations (i.e., allows the developer to specify a time frame to perform operations on the data that flows in that time window). There is sliding interval in the window, which is the time interval of updating the window.
- Provides an API in Scala, Java, and Python.
- For a stream of weblogs, if you want to get alerts within seconds- Spark Streaming is helpful.

Spark SQL

- Spark SQL provides functions for manipulating large sets of distributed, structured data using a SQL subset supported by Spark and HQL.
- It is used for reading and writing data to and from JSON files, Parquet files, Avro files, RDBMSs, Hive, etc.
- Using Spark SQL, you can 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 using HSpark
- Run unmodified Hive queries on existing data.
- Connect through JDBC or ODBC using Thrift server.

Data Visualization

- Big Data is made of numbers & 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 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 where to place which product
  - Predict sales volumes

Project Parts Details

- **PART 1.** [5] **Create your own project for Spark Streaming.**
  - ✓ Remember, it should be interesting and useful.
  - ✓ Provide detailed instructions.

- **PART 2.** [5] **Create your own project using Spark SQL and HBase/Hive together.**
  - ✓ Provide detailed instructions.

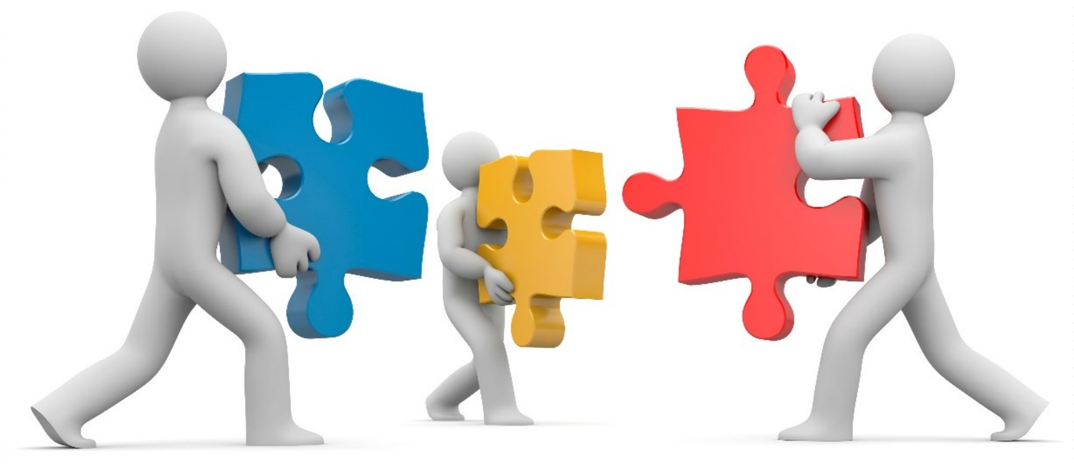
- **PART 3.** [4] **For any of the parts 1 or 2 above, show the results using any data visualization tools like** Tableau, Jupyter, Plotly, Kibana, Zeppelin.

- **PART 4.** [4] **Do some research and create a simple demo project for any one of these tools:** Presto, Impala, Phoenix, Storm, Kafka

- **PART 5.** [2] **Online Presentation of all the above 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 source files
- Shell script files for each project part wherein 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.
  - Presentation ppt can serve as Readme file if it has all the details
- Submit a .zip file of all the above-mentioned documents.