SANTANDER
PRODUCT RECOMMENDATION

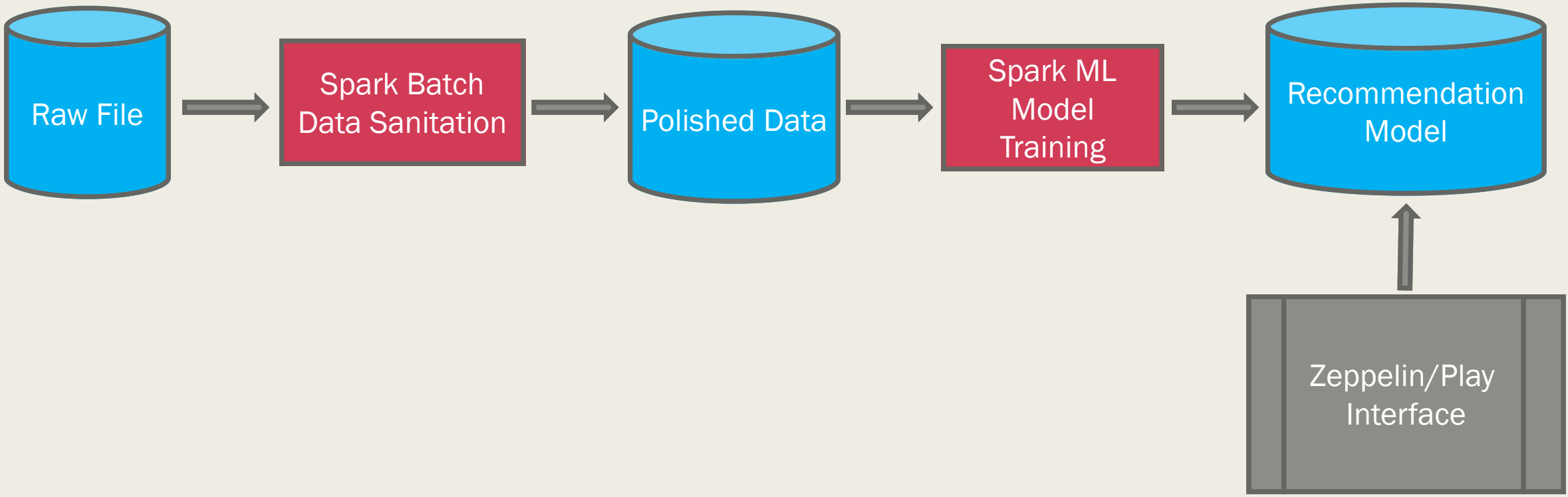
COURSE: CSYE7200 BIG DATA ENGINEERING WITH SCALA
PROFESSOR: ROBIN HILLYARD

Team 7
Arpit Rawat
Vaishali Lambe
Nishant Gandhi

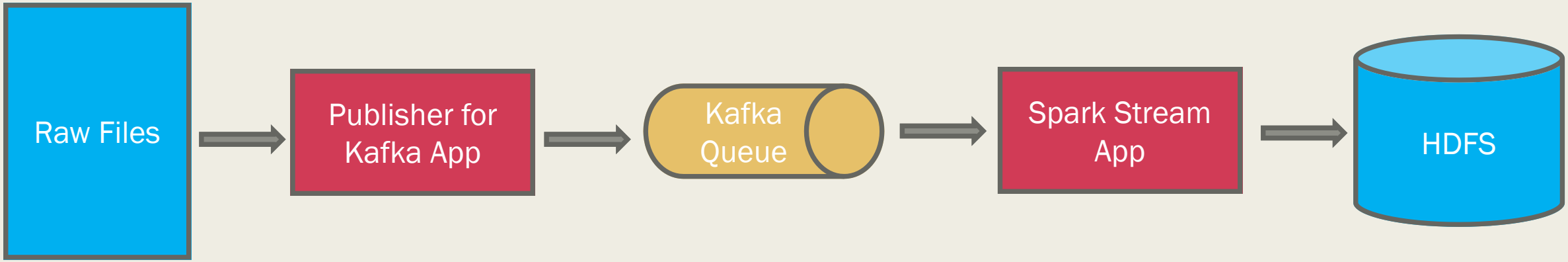
Use cases

1. Recommend New Customer the top-rated products of the bank
2. Predict which products their existing customers will use in the next month based on their past behavior and that of similar customers of the bank.

Execution Pipeline



ETL Pipeline(Additional)



Data Source: https://www.kaggle.com/c/santander-product-recommendation/data

Data Size: ~2.3GB [~930K rows]

Milestones/Sprints:

Sprint#	Timeline	Tasks
Sprint 1	03/19/2018 - 03/25/2018	Environment Set Up, Data Cleaning, Data Visualization, Unit Test Cases, Testing
Sprint 2	03/26/2018 - 04/02/2018	ML Spark, Data Modelling, Integration with pipeline, Testing
Sprint 3	04/03/2018 - 04/10/2018	Model fitting and cross validation, Testing, Accuracy, Optimization, UI framework integration
Sprint 4	04/10/2018 - 04/15/2018	Re-Testing and finishing, Preparation of Final Presentation

What you program in Scala:

- Data Extraction, Data Cleaning [IntelliJ IDEA]
- Data Visualization[Zeppelin]
- Data Modelling [Spark]
- User Interface [Play Framework]

Code Repository Location:

https://github.com/vaishalilambe/Team7_Santander_Product_Recommendation

Acceptance Criteria:

- Predictive model accuracy will be approx. 60-70%
- Running any unit test case will take approximately less than 5-10sec
- UI efficiency to get any recommendation of a banking product for pre-defined customer will be approx. 5-10sec

Goals of the Project:

- Learn and enhance skills on Scala programming, Apache Spark, Data Analysis and Apache Kafka [Kafka is additional]
- Achieve the stated acceptance criteria [refer prior slide# 7]
- Meet deadlines of each sprint and activities associated with it
- Deliver project on time without failure
- Collaborative learning and knowledge sharing, to have good team work
- Sharpen presentation skills in a group

Thank you!