Learning Program for Visa - Effective Spark - V3 - Final

Updated Feb 01, 2020

This is the draft 2 - updated from previous draft, incorporating feedback from reviewers.

# Some Questions?

* What language are we using for Spark? Scala / Java / Python / R ?
  + Prefer python, allow both
* What dataset we might use for labs? the following are some options.
  + Public datasets (pretty small, large datasets are hard to come by)
  + Synthetic data (generated) to resemble data used in Visa
  + [See this document for some ideas](https://docs.google.com/document/d/1oZQdcR_qNl-yvuV74A6rGRAWAwIv8na1s-pIhs07oEI/edit) on generating data using ISO format
* Is it possible to have access to a large scale Spark cluster to develop and validate labs?
  + TBD : Visa
* What learning services are available to participants
  + Udemy, linkedin, o'reilly
* How do we leverage internal Visa resources ? code , presentation
  + JP says, currently we have the following available
    - collating small files
    - Notebook as a service (python connected to real data) for data scientists

# Format for each lesson

* Knowledge check before
* Reading materials (book chapters, blogs ..etc)
* Videos to watch (Pluralsight, youtube ..etc)
* Knowledge check after
* Labs (easy, medium, hard)

# How do we measure learning?

* Before / After quiz
  + We will do it per session
  + This allows a student to measure they proficiency for each section
  + And they can choose to take the session or skip it based on their self assessment
* Any other metrics?
  + Internal to Visa
  + This is hard to quantify at this point. May be later

# Section 1 : Spark Basics

This section serves as a quick introduction to Spark

## Spark introduction (essential coverage only)

* Architecture
* Overview of Spark components
* Use cases
* Knowledge check
* Labs
  + Getting access to Spark cluster
  + Or setting up Spark on a local machine to experiment

## Dataframes

* Dataframes introduction
* Dataframes APIs
* Knowledge check
* Labs
  + Loading and analyzing data

## Spark SQL

* Introduction to Spark SQL
* Spark SQL vs Hive SQL
* Architecture
* SQL features
* Catalyst optimizer (TBD may be move to intermediate)
* Knowledge check
* Labs
  + Loading and analyzing data using Spark SQL
  + Measuring optimizer
  + Optimizing sql query performance iteratively

## Spark and Hive

* Assume they know basic hive, hiveql
* Right now HiveQL is used. Goal is to move to SparkSQL
* Accessing Hive tables from Spark
* Understanding spark catalog
* Knowledge check
* Labs
  + Querying hive tables from Spark

# Section 2 : Intermediate Spark

## Spark and HDFS

* HDFS, Spark and localized processing
* Partitions and distributed processing
* Understanding block size and its implications
* Dealing with small files
* Knowledge check
* Labs:
  + Generating large scale data and uploading into HDFS
  + How to deal with too many small files
    - How to automatically collate small files (during streaming for example)
  + Accessing HDFS files from Spark
  + Experimenting with file block sizes and Spark processing

## Spark and YARN

* YARN overview
* Understanding Spark and YARN interaction
* Client / cluster mode
* YARN settings that affect Spark Jobs
* Knowledge check
* Labs:
  + Submitting a spark job to YARN

## Effective Caching in Spark

* Understanding Spark caching
* Caching dataframes
* Caching SQL tables
* Measuring performance of cached data
* Knowledge check
* Labs:
  + Measuring Cache performance

## Effective Joins in Spark

* Understanding join mechanism
* Understanding shuffles
* Techniques for improving join performance
* Knowledge check
* Labs:
  + Joins

## Restartable Pipelines

* Thinking in omnipotent steps
* Best practices applied to Spark pipelines
* Labs
  + Constructing a restartable pipeline

## Effective Data Formats in Spark / Hadoop

* Understanding various dataformats (AVRO, JSON, Parquet, .etc)
* -Effective data architecture using these formats:
  + Storage optimizations (compression, partitioning) for performance
  + Creating large files using compaction, “reconciliation” etc and the impact on applications
* Converting from one data format to another using Spark APIs
* Knowledge check
* Labs
  + Experimenting with various data formats
  + Measuring query performance on various data formats

## Kafka Streaming With Spark

* Structured streaming with Spark and Kafka
* Setup a test environment with Kafka + Spark
* Running queries on kafka streams
* Knowledge check
* Labs
  + Kafka + Spark setup
  + Kafka + Spark labs