*Hortonworks Technical Exercise*

As discussed, here is an exercise we give candidates to complete on their own time and then present to Hortonworkers. Please reach out as soon as you complete the exercise below. Candidates may submit up to $100 in expenses back to the Hortonworks Technical Recruiter for reimbursement after completion of this exercise. Let me know if you have questions:

Task 1 – HDP Cluster Installation

Set up a 3-node (at least) cluster of the latest Hortonworks Data Platform via Ambari installation. You'll need to evaluate which components are actually required based on the rest of the exercise.

Task 2 – Data Pipelines in Hive

1. Install the [MySQL Employees sample database](https://dev.mysql.com/doc/employee/en/employees-installation.html" \t "_blank) (available from [Employees DB on Launchpad](https://launchpad.net/test-db/" \t "_blank)) within Hive’s MySQL instance
2. Perform a full import of the ‘employees’ and ‘salaries’ tables into HDP
3. Create Hive tables out of them using a suitable Storage Format
4. Perform some cleansing using either Pig, Hive or other tool set
5. The ‘to\_date’ column overlaps with the ‘from\_date’ column in the ‘salaries’ table.  This results in the employee having two salary records on the day of the ‘to\_date’ column. Fix it by decrementing the ‘to\_date’ column by one day (e.g. from ‘1987-06-18’ to ‘1987-06-17’) to make each salary record exclusive.
6. The first salary record for an employee should reflect the day they joined the company.  However, the ‘employees.hire\_date’ column doesn’t always reflect this.  Clean the data by replacing the ‘employees.hire\_date’ column with the first salary record for an employee.
7. Determine which employee lasted less than two weeks in the job in May 1985?

*\*For Task 3, choose one of the options below*

Task 3 – Option 1: Streaming Architecture with Hive

Hive

1. Store the twitter data in the attached file called sample\_twitter\_data in HDFS. The data is in json format and should not be altered.
2. Once the data is in HDFS, create an hcat/hive schema to be able to answer the following question: What are all the tweets by the twitter user "Aimee\_Cottle"? You will need to provide the query that answers this question.

*Hint: there are multiple ways to do this, the preferred method involves org.apache.hcatalog.data.JsonSerDe - if that doesn't work search for Json serde's in the www - there are some you can compile from source to get it to work*

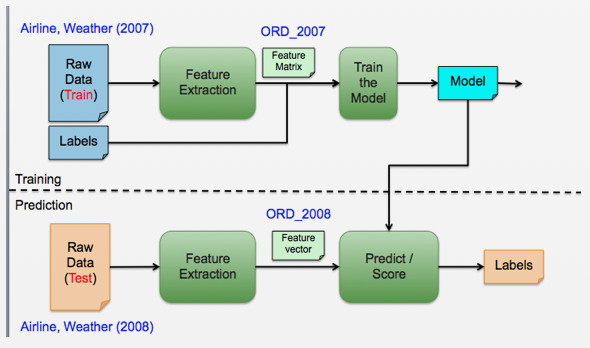
Streaming

1. Implement a [storm](http://hortonworks.com/hadoop/storm/) topology that streams in tweets (<https://dev.twitter.com/streaming/overview>), does some interesting analytics in real-time on the tweets, and then persists into HDFS.

Task 3 – Option 2: Data Science using Spark, PySpark and IPython

We suggest tackling this task using either the [Zeppelin Tech Preview](http://hortonworks.com/hadoop-tutorial/apache-zeppelin/) or the [Ipython Notebook](http://ipython.org/notebook.html) (in which case you should create the model in scala using the [ISpark plugin for IPython Notebook](https://github.com/tribbloid/ISpark))

1. Download the following datasets:
   1. [Airline data](https://www.dropbox.com/s/c7pn0qseqfhz7pw/airline.tar.gz?dl=0)
   2. [Weather Data](https://www.dropbox.com/s/rb47vyjaixexzn3/weather.tar.gz?dl=0)
2. Goal: Predict delay (delayTime > 15 mins) in flights – limit your analysis to ORD for simplicity
3. Use Spark and [MLLib](https://spark.apache.org/mllib/)'s logistic regression to perform this prediction.
4. Your model should be written in scala
5. You can use the following pig script to join the datasets (the preprocessing) and create the feature matrix: <https://www.dropbox.com/s/op00i9i19dydyzf/preprocess.pig?dl=0> <https://www.dropbox.com/s/j1mx1nawnw1nt5h/util.py?dl=0>
6. The following flow should help with what is expected:



Task 4 – Solution Design

*Given the example below please provide a high level Solution Architecture using the HDP stack in order to satisfy the customer’s requirements. This should be a presentation containing a high level architecture with high-level details and explanations for the solution proposed. Sizing of the cluster (number of nodes) would be considered a bonus.*

Mortgage Processing Express LLC

We are a company that processes mortgage documents for all major banks and corporations across the United States. We provide our clients with [eDiscovery](https://en.wikipedia.org/wiki/Electronic_discovery) of the documents as well as detailed analysis of any dollar amounts within the documents themselves. Currently our MPP infrastructure is unable to handle the velocity of mortgage documents flowing in and we are turning to Hadoop. We’ve listed the technical explanation of all our sources and the data points we extract during eDiscovery below.

3 sources of mortgage documents flow into our application

1. Data landing into Amazon S3.   
   Anywhere from 100 million to 300 million documents on average land here per day. This is more of a batch environment and can be processed daily.
2. Relational database systems - Oracle, SQL Server.   
   The data in these sources is stored in blobs and is surfaced through our web applications. We see approximately five million transactions a day on average - peaking up to ten million. These include new mortgage documents and updates. The processing of this data should happen in less than 30 minute intervals.
3. Our ActiveMQ system   
   Here we track real-time updates and allow our customers to post documents (due to their clients requiring second response updates on modified documents). We have well over ten million transactions a day here with new documents and updates. The processing of these mortgage documents must be performed in real time as our clients demand constant real-time updates.

The data types for our documents range from MS word, Excel and PDF documents in their raw text form, to typed image documents as well. Basically we need to process and extract raw text from the documents as well perform some form of image parsing from the image-based documents. Another level of complexity is introduced as certain image documents have hand written notes on them which we also need to parse and extract content for.

Document extraction requirements:

* Ability to consume all 3 sources of documents from S3, RDBMS systems and MQ systems. The time to submission to processing the documents should be more efficient than our current systems which are time constrained today.
* Extract raw text from all documents (Raw, text, Images, hand written note images) and make available for real-time search and surfaced through Java APIs and REST for our web applications.
* Extract metadata (creation time, modified time, etc.)
* During processing mortgage dollar amounts (e.g. “$15,000”) should be extracted separately from the raw text and made available for our BI reporting through some form of SQL interface
* The documents themselves should also be available for retrieval during real-time search

Task 5 – Presentation

Prepare a short presentation (4-6 slides or so) on an initial pitch of "Why Hadoop for this use case?” and “Why Hortonworks for Hadoop?” based on what you can find on our website and the general www. As an alternative, you can also choose to present on a subject matter you are already familiar with. This can be existing sales pitches you have performed in current or previous roles as well as sales-focused presentations around projects you have performed in the past.

Also prepare to speak to how you tackled this assignment and what you learned as you demo the solution. You will be asked to present in person (if logistics allow) assuming that I am the potential customer evaluating Hadoop. The presentation/demonstration is expected to be one hour, including Q&A.