**Analyze data sets for applicable strategy**

Your company has recently bagged a large assignment from a US-based customer that is into training and development. The larger outcome deals with launching a suite of educational and skill development programs to consumers across the globe. As part of the project, the customer wants your company to analyze a series of data sets to arrive at a prudent product mix, product positioning, and marketing strategy that will be applicable for at least a decade. The following are the expected outcomes of this project:

* -  Identify motivators for continuous adult education.
* -  Identify occupations poised for growth & decline over the next 10 years.
* -  Identify regions that have potential for growth across potential industries.
* -  Categorize financial capacity of consumers across regions and demographics.
* -  Identify major gender and geographic attributes for education.
* -  Analyze the education expenditure and related parameters across the globe.
* -  Analyze the strength of the financial sector in target markets and participation of the population  in the financial sectors.  In line with the expected outcomes above, the whole project is divided into 7 subprojects, each involving its own data set. The customer has also provided URLs for all the data sets. Your project manager has assigned you the responsibility of completing all the subprojects within the defined timeframe. The details of the subprojects and the scope of each subproject are listed in the sections below.

**Motivators for Continuous Adult Education**

The first subproject deals with analyzing educational data sets to identify motivators for continuous learning. The scope of this subproject is listed below:

* -  Analyze the entire data set and arrive at all designations where the annual salary is above $60,000.
* -  Arrive at the education level requirement for the above salary range.
* -  Arrive at the designations where the education level is “High school diploma or equivalent” or  less.
* -  Analyze the average replacement needs for the salary range in step 1 for the year 2010-2012.  You can download the required data set in XLS format from the URL below:  http://www.bls.gov/emp/ep\_table\_111.htm
* The total time provided by your manager to complete this task is 8 hours.  Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.

**Occupations Poised for Growth**

The next subproject is about identifying occupations poised for growth in over a decade. The scope of this subproject is listed below:

* -  List all occupations that are expected to grow by over 20% by 2022.
* -  List the education levels needed for such occupations.
* -  Identify and list all occupations that are poised for decline of more than 10% by 2022.  You can download the required data set in XLS format from the URL below:  http://www.bls.gov/emp/ep\_table\_111.htm
* The total time allocated by the project manager to perform this task is 4 hours.  Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.
* **Region-Wise Job Opportunities** for Top Occupations  As the next step, you have to identify the regions (states) in which job opportunities exist for the top occupations identified in the previous subproject. The scope of this subproject is listed below:
* -  List all states where growth in job opportunities for the jobs listed in the previous subproject is more than 30%.
* -  List all occupations where the annual median wage is more than 50% for the job opportunities listed in the previous scope of this project.

You can download the required data set in XLS format from the URL below:

http://www.bls.gov/oes/tables.htm

(Download the STATE (XLS) file under the May 2013 Section)

The total time allocated to perform this task is 4 hours by the project manager.

Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.

**Financial Capacity of Consumers**

As the next step, you have to identify the financial capacity of the regions identified in the previous subproject. The scope of this subproject is listed below:

* -  Identify the industries and regions from the previous project.
* -  For the regions and industries listed in the previous step, identify the average weekly wages.  You can download the required data set in XLS format from the URL below:  http://www.bls.gov/web/cewqtr.supp.toc.htm  (Download the data available in the Archive of Excel workbook’s (.xls's) section) The total time allocated to perform this task is 4 hours by the project manager.  Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.  Gender and Geographic Attributes for Education  This subproject deals with identifying the gender factor in higher education aligned to the geographies that make an impact. The goal of this subproject is to identify special programs aimed at women across

the globe.

The following is the scope of this subproject:

* -  Load the data in a relational database like MySQL.
* -  Import the data into Hive.
* -  Identify the countries where % of female graduates is less than 30%.
* -  List the average increase in female education in the U.S. from the year 2000.
* -  List the % of change in male employment from the year 2000.
* -  List the % of change in female employment from the year 2000.  You can download the required data set in XLS format from the URL below: © Copyright 2015, Simplilearn. All rights reserved.

http://data.worldbank.org/data-catalog/gender-statistics

(Download the Excel data available in the resources section) The total time allocated by the project manager to perform this task is 6 hours. Hint: Use the techniques taught in lesson 11 to use Sqoop.

**Education Expenditure and Related Analysis Across the Globe**

To ensure the customer is able to target appropriate geographies beyond the U.S. for skill development, your next task is to analyze a UN data set to arrive at the economies that spend the highest and least on education and compare the same with the U.S. The following is the scope of this subproject:

* -  Identify the top 5 countries based on education expenditure (current US$) since 2000.
* -  List the gross savings (% of GNI) for the top 5 countries since 2000.
* -  Export the result to HBase.
* -  Add currency for each of the 5 countries in the HBase table.  You can download the required data set in XLS format from the URL below:  http://data.worldbank.org/data-catalog/world-development-indicators
* (Download the Excel data available in the resources section) The total time allocated to perform this task is 6 hours by the project manager.  Hint: Use the techniques taught in lesson 10 to use HBase. Export data from Pig/Hive using techniques mentioned in Lesson 07/08.

**Financial Sector Strength in the Identified Markets**

As the last step toward completing the entire project, your customer wants you to analyze the strength of the financial sector and the participation of the population in financial activities in the economies identified in the previous subproject. The following is the scope of this subproject:

* -  Identify the central bank assets to GDP (%) growth percentage since 2000 in the economies identified in the previous subproject.
* -  Analyze the bank return on assets (%, after tax) change % in the identified economies since 2000.
* -  Identify the average domestic credit to private sector (% of GDP) since 2000 in the identified economies.
* -  Calculate the average stock market capitalization to GDP (%) since 2000 in the identified economies.
* -  Identify the change % in bank accounts per 1,000 adults in the identified economies since 2000.
* -  Calculate the average % saved at a financial institution in the past year (% age 15+).
* -  Write the above steps as a script, run and schedule using a workflow in Oozie.  You can download the required data set in XLS format from the URL below:  http://data.worldbank.org/data-catalog/global-financial-development  (Download the Excel data available in the resources section) The total time allocated to perform this task is 8 hours by the project manager.  Hint: Use the techniques taught in lesson 07/08 to run scripts in batch mode. Schedule using techniques shown in demo of lesson 12.

**Analyze data sets for applicable strategy**

Your company has recently bagged a large assignment from a US-based customer that is into training and development. The larger outcome deals with launching a suite of educational and skill development programs to consumers across the globe. As part of the project, the customer wants your company to analyze a series of data sets to arrive at a prudent product mix, product positioning, and marketing strategy that will be applicable for at least a decade. The following are the expected outcomes of this project:

* -  Identify motivators for continuous adult education.
* -  Identify occupations poised for growth & decline over the next 10 years.
* -  Identify regions that have potential for growth across potential industries.
* -  Categorize financial capacity of consumers across regions and demographics.
* -  Identify major gender and geographic attributes for education.
* -  Analyze the education expenditure and related parameters across the globe.
* -  Analyze the strength of the financial sector in target markets and participation of the population  in the financial sectors.  In line with the expected outcomes above, the whole project is divided into 7 subprojects, each involving its own data set. The customer has also provided URLs for all the data sets. Your project manager has assigned you the responsibility of completing all the subprojects within the defined timeframe. The details of the subprojects and the scope of each subproject are listed in the sections below.

**Motivators for Continuous Adult Education**

The first subproject deals with analyzing educational data sets to identify motivators for continuous learning. The scope of this subproject is listed below:

* -  Analyze the entire data set and arrive at all designations where the annual salary is above $60,000.
* -  Arrive at the education level requirement for the above salary range.
* -  Arrive at the designations where the education level is “High school diploma or equivalent” or  less.
* -  Analyze the average replacement needs for the salary range in step 1 for the year 2010-2012.  You can download the required data set in XLS format from the URL below:  http://www.bls.gov/emp/ep\_table\_111.htm  The total time provided by your manager to complete this task is 8 hours.  Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.

Occupations Poised for Growth

The next subproject is about identifying occupations poised for growth in over a decade. The scope of this subproject is listed below:

* -  List all occupations that are expected to grow by over 20% by 2022.
* -  List the education levels needed for such occupations.
* -  Identify and list all occupations that are poised for decline of more than 10% by 2022.  You can download the required data set in XLS format from the URL below:  http://www.bls.gov/emp/ep\_table\_111.htm  The total time allocated by the project manager to perform this task is 4 hours.  Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.  Region-Wise Job Opportunities for Top Occupations  As the next step, you have to identify the regions (states) in which job opportunities exist for the top occupations identified in the previous subproject. The scope of this subproject is listed below:
* -  List all states where growth in job opportunities for the jobs listed in the previous subproject is more than 30%.
* -  List all occupations where the annual median wage is more than 50% for the job opportunities listed in the previous scope of this project.

You can download the required data set in XLS format from the URL below:

http://www.bls.gov/oes/tables.htm

(Download the STATE (XLS) file under the May 2013 Section)

The total time allocated to perform this task is 4 hours by the project manager.

Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.

**Financial Capacity of Consumers**

As the next step, you have to identify the financial capacity of the regions identified in the previous subproject. The scope of this subproject is listed below:

* -  Identify the industries and regions from the previous project.
* -  For the regions and industries listed in the previous step, identify the average weekly wages.  You can download the required data set in XLS format from the URL below:  http://www.bls.gov/web/cewqtr.supp.toc.htm
* (Download the data available in the Archive of Excel workbook’s (.xls's) section) The total time allocated to perform this task is 4 hours by the project manager.  Hint: Use the techniques taught in lesson 3 to upload the data in HDFS. Hive/Pig is the most recommended programming tool to accomplish this task.  Gender and Geographic Attributes for Education  This subproject deals with identifying the gender factor in higher education aligned to the geographies that make an impact. The goal of this subproject is to identify special programs aimed at women across

the globe.

The following is the scope of this subproject:

* -  Load the data in a relational database like MySQL.
* -  Import the data into Hive.
* -  Identify the countries where % of female graduates is less than 30%.
* -  List the average increase in female education in the U.S. from the year 2000.
* -  List the % of change in male employment from the year 2000.
* -  List the % of change in female employment from the year 2000.  You can download the required data set in XLS format from the URL below: © Copyright 2015, Simplilearn. All rights reserved.

http://data.worldbank.org/data-catalog/gender-statistics

(Download the Excel data available in the resources section) The total time allocated by the project manager to perform this task is 6 hours. Hint: Use the techniques taught in lesson 11 to use Sqoop.

Education Expenditure and Related Analysis Across the Globe

To ensure the customer is able to target appropriate geographies beyond the U.S. for skill development, your next task is to analyze a UN data set to arrive at the economies that spend the highest and least on education and compare the same with the U.S. The following is the scope of this subproject:

* -  Identify the top 5 countries based on education expenditure (current US$) since 2000.
* -  List the gross savings (% of GNI) for the top 5 countries since 2000.
* -  Export the result to HBase.
* -  Add currency for each of the 5 countries in the HBase table.  You can download the required data set in XLS format from the URL below:  http://data.worldbank.org/data-catalog/world-development-indicators
* (Download the Excel data available in the resources section) The total time allocated to perform this task is 6 hours by the project manager.  Hint: Use the techniques taught in lesson 10 to use HBase.
* Export data from Pig/Hive using techniques mentioned in Lesson 07/08.

**Financial Sector Strength in the Identified Markets**

As the last step toward completing the entire project, your customer wants you to analyze the strength of the financial sector and the participation of the population in financial activities in the economies identified in the previous subproject. The following is the scope of this subproject:

* -  Identify the central bank assets to GDP (%) growth percentage since 2000 in the economies identified in the previous subproject.
* -  Analyze the bank return on assets (%, after tax) change % in the identified economies since 2000.
* -  Identify the average domestic credit to private sector (% of GDP) since 2000 in the identified economies.
* -  Calculate the average stock market capitalization to GDP (%) since 2000 in the identified economies.
* -  Identify the change % in bank accounts per 1,000 adults in the identified economies since 2000.
* -  Calculate the average % saved at a financial institution in the past year (% age 15+).
* -  Write the above steps as a script, run and schedule using a workflow in Oozie.  You can download the required data set in XLS format from the URL below:  http://data.worldbank.org/data-catalog/global-financial-development  (Download the Excel data available in the resources section) The total time allocated to perform this task is 8 hours by the project manager.  Hint: Use the techniques taught in lesson 07/08 to run scripts in batch mode. Schedule using techniques shown in demo of lesson 12.