**General Autodiagnostic (AD) Workshop**

QM69 – previous nomenclature of AD

**1. Basic AD Guidelines**

1.1. Version Control: Bitbucket

- It is used as an online repository for source codes and development projects in Micron

- Github is very similar to Bitbucket, but the former is a more open source repository for both public and private use. On the other hand, bitbucket is not as open as Github in terms of source codes; it appeals more to enterprises who are developing proprietary source codes and keeping most of their trade secrets.

- Type “http://stash/” in windows explorer to access Bitbucket easily

- It is highly recommended to develop the habit of storing codes at bitbucket from the onset, to prevent any potential loss of data

- Before writing and storing any code on Bitbucket, please create a README.md as a readable guide on how the algorithm functions

- GitKraken is also recommended to be used alongside Bitbucket; it is a graphical user interface (GUI) that is useful in keeping the repository organized. GitKraken is particularly useful in visualizing and tracking any changes in the repository

1.2. Documentation: Confluence

- Similar to Wikipedia, it is a centralized destination for the technical information relevant to the DS team

- Many useful APIs and projects can be found under the “Data Source Repository” section

- Data Science APIs are meant to be abstracted modules for data access, transformation, analytics and actions. These APIs reduce duplication of same effort by various teams for similar functionalities

- Example: accessing probe bins in Spark environment. These modules are written in a way to increase reusability and reproducibility for data science applications.

- Each project's main page has *basic information* that are frequently referenced by DS team members and IT EAD & support teams

1.3. Project Tracking: JIRA

- It is a project management tool that supports project updates and other agile methodology. For example, it can be used to track updates and status reports on the project at hand, and provide issue and bug trackings

- Priorities can be set on the projects to further distinguish their order of importance (P1, P2, P3, etc.)

1.4. Python Environment Management: Anaconda

- It is a free and open source distribution of Python and R programming languages for data science applications

-It is a desktop graphical user interface (GUI) included in Anaconda distribution that allows to launch applications and manage conda packages, environments and channels without using command-line commands

- A wide variety of machine learning libraries are available in Anaconda

- An advantage of using Anaconda is that it can be used for handling different versions of Python packages

- Applications downloaded may not always be maintained at the latest version, so it might be good to set up another environment in which it can work properly

1.5.1. IDE: Python: Spyder

- It is an open source integrated development environment (IDE) for scientific programming in Python

- It comes along with the Anaconda distribution package

- Several data science libraries are included, such as NumPy, SciPy, Matplotlib, and IPython

1.5.2. IDE: Python: Pycharm

- It is another IDE besides Spyder. It provides code analysis, a graphical debugger, an integrated unit tester, integration with version control systems (VCSes) and supports web development with Django

1.5.3. IDE: Java: IntelliJ

- It is an IDE for Java

1.5.4. IDE: Web: Visual Studio Code and Bracket

- VS code and Brackets are source code editors with a primary focus on web development

**2. Processing of Raw Data and Database**

- Raw data are created and stored by the end-users in Micron (e.g. RDA, PEE) in their own source systems at the very onset

- Data is required to be extracted and copied to a data warehouse via a means known as extract, transform and load (ETL)

- ETL are three database functions that are combined into one tool to migrate data from one database to another. It converts databases from one format or type to another suitable format

- the EAD team will apply ETL on raw data then ingest into GDW OR ingest into GDW directly, it depends on the data format (e.g. for unstructured data, we apply ETL firstly)

- to finish data preparation, it involves several concepts as follows:

- Global data warehouse (GDW): HDFS vs Teradata

- [Hadoop Distributed File System](https://www.edureka.co/blog/hdfs-tutorial) (HDFS) is the core component or the backbone of Hadoop Ecosystem. It makes it feasible to store different types of large data sets (ie. structured, unstructured and semi structured data)

- HDFS creates a level of abstraction over the resources, from where we can see the whole HDFS as a single unit

- HDFS has two core components, i.e. **NameNode and DataNode**

- HDFS is open source whereas Teradata is commercial; HDFS is backed by a growing community fixing bugs and making improvements on a consistent basis whereas Teradata (RDBMS) is capable of executing complex queries (ie. more specialized)

- Hive and HBase are 2 hadoop based big data technologies that serve different purposes: they are both data stores for storing unstructured data

- Hbase is a low latency NoSQL database used for real-time data streaming. It does not provide random access capabilities, is highly scalable to produce high throughput, available, and consistent

- The best way to use HBase is to make Hadoop the repository for static data and Hbase the data store for data that is going to change in real-time after some processing. It should be used when there is a large amount of data and data model schema is sparse

- On the other hand, Hive is a mapreduce based SQL engine that runs on top of Hadoop (ie. batch processing of big data. It is designed for SQL savvy people to run mapreduce jobs through SQL like queries

- Hive allows developers to impose a logical relational schema on various file formats and physical storage mechanisms within or outside the Hadoop cluster; more productive than writing mapreduce directly (5 lines of HiveQL >= 100 lines of Java)

- RDBMS professionals love Hive as they can simply map HDFS files to Hive tables and query the data. However, for big data applications that require complex and fine grained processing, Hadoop mapreduce is the best choice

- NEXUS is the only tool that works on every system including Hadoop. It allows users to see tables and views from all accessible systems and join data across many different platforms

- its Super Join Builder allows for the ERwin Logical Model to be loaded, and then Nexus shows tables and views visually.  It then guides users to show what joins to what

- Dialects and differences exist within SQL that result in many types of SQL, because of various reasons such as the difference in platforms, standards specifications

- MSSQL server is a RDBMS developed by Microsoft; it is built for the basic function of storing and retrieving data as required by other applications; it can be run either on the same computer or on another across a network

- SQL databases are primarily classified as RDBMS, whereas NoSQL database are primarily called non-relational or distributed database; the former is a table based database while NoSQL databases are document based, key-value pairs. SQL databases have predefined schema whereas NoSQL databases have dynamic schema for unstructured data

- Self-learning for MSSQL:

* create table with schema
* ingest data into MSSQL
* select data from table and to do join

**3. Extraction of Data from One or Multiple Data Source**

3.1. Bash

- It is a Unix shell and command language for most Linux distributions; it is a command processor that typically runs in a text window where the user types commands that cause actions.

- Login into server: we normally use MobaXterm (recommended) or Putty/MPutty

- Basic commands:

* Basic Terminal Shortcuts
* Basic Terminal Navigation
* Basic File manipulation
* Research Files
* Extract, sort and filter data
* Time Setting

3.2. Ambari

- The Apache Ambari project is aimed at making Hadoop management simpler by developing software for provisioning, managing, and monitoring Apache Hadoop clusters. Ambari provides an intuitive, easy-to-use Hadoop management web UI backed by its RESTful APIs.

- Hive view and HDFS file view

***NOTE:***

For handling data in HDFS, there are several methods as follows,

* **Bash**

Hadoop fs -text

Hive –hiveconf

* **Java**

MapReduce

* **Spark/PySpark**

Sql

* **Python**

mu\_hbasethrift

3.3. Teradata Studio

- Teradata Studio is a consolidated DBA and query tool for the Teradata Database. It allows you to perform common DBA administrative functions, create and submit queries, and it includes a self-service smart loader for Teradata Data Lab.

***NOTE:***

Several ways for Teradata:

* **Bteq (Basic Teradata Query)**

BTEQ was the first utility and query tool for Teradata. BTEQ can be used as a Query tool, to load data a row at a time into Teradata and to export data off of Teradata a row at a time.

* **Python**Teradata (package)
* **Java**
* JDBC: Java JDBC is a Java API to connect and execute query with data base. JDBC API uses jdbc drivers to connect with the database.
* There are 5 steps to connect any java application with the database using JDBC. These steps are as follows:

1. Register the driver class
2. Creating connection
3. Creating statement
4. Executing queries
5. Closing connection

3.4. MS sql studio

***NOTE:***

Several ways as follows:

* **Python**

Pymssql:

* A simple database interface for [Python](http://www.python.org/) that builds on top of [FreeTDS](http://www.freetds.org/) to provide a Python DB-API ([PEP-249](http://www.python.org/dev/peps/pep-0249/)) interface to [Microsoft SQL Server](http://www.microsoft.com/sqlserver/).

[Step 1: Configure development environment for pymssql Python development](https://docs.microsoft.com/en-us/sql/connect/python/pymssql/step-1-configure-development-environment-for-pymssql-python-development?view=sql-server-2017)

[Step 2: Create a SQL database for pymssql Python development](https://docs.microsoft.com/en-us/sql/connect/python/pymssql/step-2-create-a-sql-database-for-pymssql-python-development?view=sql-server-2017)

[Step 3: Proof of concept connecting to SQL using pymssql](https://docs.microsoft.com/en-us/sql/connect/python/pymssql/step-3-proof-of-concept-connecting-to-sql-using-pymssql?view=sql-server-2017)

More information: <https://docs.microsoft.com/en-us/sql/connect/python/pymssql/python-sql-driver-pymssql?view=sql-server-2017>

* **Java**jdbc
* **Digger**
* **Yield Cube/MICA**Typically Yield Cube is for PEE because they don’t have access to HDFS. But Yield Cube is a fast way for us to query some dataset or specified data.

**4. Analysis of Data**

- Analysis of data can be done on a variety of platforms, such as Spyder, R, Spark, etc.

4.1. Spyder

- It is an interactive development environment (IDE) for Python with advanced editing, interactive testing, debugging and introspection features

- Its GUI is very similar to Matlab; it has a numerical computing environment and popular Python libraries such as NumPy, SciPy and matplotlib

4.2. Jupyter Notebook

- On the other hand, the Jupyter Notebook is a server-client application that allows editing and running notebook documents via a web browser, and it allows for displaying/editing/running notebook documents

4.3. Scikit-learn

- It is a free software machine learning library for Python, which includes supervised and unsupervised learning methods like classification, regression, support vector machines, random forests, gradient boosting and k means

4.4. R

- On the other hand, R is a programming language with a slant towards statistical computing and graphics. It has an extensive library of machine learning libraries but it smaller in data processing scale as compared to Python or Java

4.5. ANOVA

- The Analysis of Variance (ANOVA) test is statistical method of finding out if the survey or experiment results are significant (ie. testing of null vs alternate hypothesis

4.6. MapReduce

- It is a processing technique and a program model for distributed computing based on Java

- It consists of 2 functions: map and reduce: the map function always run first, and it is typically used to filter, transform and parse data. The reduce function is optional, and its normally used to summarize data from the map function

- Just like HDFS, MapReduce has master and slave nodes as well. The master node runs master daemons to accept jobs, and monitor and distribute work

- The slave nodes run slave daemons to start tasks, do actual work and report status back to master daemons

**5. Save Analysis Result**

- Once analysis has been done, it is important to store the results on appropriate softwares for subsequent deployment

-Multiple dataset from different servers may be used by project taskers and it is important to consolidate them into one

5.1. HBase

- it is an open source distributed database which would yield higher performance while being cost effective at the same time, and we do not have to worry about distributing the data for faster processing since Hadoop takes care of it

- It is also easily scalable, fault tolerant and highly available

5.2. Hive Table

- There are mainly 2 kinds of table in Hive, external and internal

- By default hive will create tables as an internal one, it is under the control of hive whereas the external table is not in control of hive; it is created outside the warehouse directory

5.3. Teradata Data Lab

- It is a separate dedicated space within a production data warehouse for agile development of new analytic queries that can combine personal, ad hoc, or temporary data with production data

- It is able to analyze, reorganize, model and prototype a sample of real-time data, and experiment without any risk to production data

- Data labs are governed by the DBA or by designated users, depending on the needs of the organization

5.4. MS SQL

- MSSQL server is a RDBMS developed by Microsoft; it is built for the basic function of storing and retrieving data as required by other applications; it can be run either on the same computer or on another across a network

- Just like HBase and Hive table, MS SQL server allows for the storage of data

**6. Display Analysis result**

6.1. Tableau

- it is a a very powerful and flexible end-to-end analytics BI tool that is capable of providing a platform for your data

- It focuses on data visualization, dashboarding and data discovery. Analysis can be conducted more easily through effective visualization of data

- The Tableau company offers a range of BI products such as Tableau Prep, Desktop, Online and Server

6.2. Python matplotlib

- Matplotlib is a Python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms, such as in Python script, IPython shells, Jupyter notebook

- It allows for generation of plots, histograms, power spectra, bar charts, errorcharts, scatterplots, etc., with a relatively few lines of code

- (<https://matplotlib.org/users/index.html> for more information)

6.3. Javascript

- It is a scripting language that enables you to create dynamically updating content, control multimedia, and animate images

- For example, it can display timely content updates, interactive maps, animated 2D/3D graphics, scrolling video jukeboxes, etc. on web pages

**7. Deployment of Model**

7.1. ActiveBatch

- It is a workload automation and job scheduling system which integrates business applications, stand-alone tasks processes, and scripts across different computing environments

- Whilst focusing on coordinating and automating on functions, it also provides insights and analytics by identifying, monitoring, and managing workflows and systems and data dependencies between them

7.2. Cron

- It is a time-based job scheduler in Unix OS; people who set up and maintain software environments use cron to schedule jobs to run periodically at fixed times, dates or intervals

- It typically automates system maintenance or administration

7.3. Nifi Apache

- It is a software project from Apache that is designed to automate the flow of data between software systems