**IBM Data Science Certificate - Started Sep 2020**

**COURSE 1 - Intro to Data Science:**

**Mod 1:** What is Data science:

* study of data, it’s a process

Fundamentals of DS: can help orgs understand enviros, analyze existing issues, discover opps

adds knowledge, helps answer questions,

* structured or unstructured
* DS can tell stories with visualizations

**Mod 2: What DS’ do**

- discover optimum solutions to existing problems.

Uber – where divers should be

TTC – best routing

Lakes – predict cyanobacterial blooms to protect public health

Ident problem – gather data, ident required tools, devel data strategy

Topics – regression, visualization, structured (ex: excel, table format), unstructured (ex: everything else.)

Regression example – Taxi rides - can calculate constant fixed fare if it knows relationship between cost of travelling by minutes or meters and for how long/far.

**Cloud** for DS: - it’s a central storage system, allows multiple entities to work with simultaneously on same data.

- allows access to open tech like apache spark and tools and libraries,

- accessible from anywhere

ex: IBM cloud, amazon web services, google cloud platform

- enchances productivity for DSs

Data scientist def – one who finds solutions to probs using analytics data with tools then comms findings. Data science is what data scientist do.

- unique blend of skills to unlock data and tell a story – this def is very inclusive of other disciplines.

be curious, ask good questions, and be comfortable with unstructured situations by making them structured.

**Module 3 - Big Data and Data Mining**

Big data - Big Data refers to the dynamic, large and disparate volumes of data being created by people, tools, and machines. It requires new, innovative, and scalable technology to collect, host, and analytically process the vast amount of data gathered in order to derive real-time business insights

that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value.”

5 V’s – Velocity – speed that data accumulates

Volume – increase or scale of amount of stored data (2.5 quintillion bytes per day currently)

Variety – structured (%20 - excel, relational databases) unstructured (%80 - not organized and everything else). – needs analyzing and organizing. data comes from diff sources and drivers

Varacity – quality and origin of data, completeness, Drivers include cost and tracability

Value – ability to turn data into value (profit, public social benefits).

Alternative tools to analyze now like Hadoop and apache spark.

Organizations using big data to transform! need support of CEO and chief info officer as well as buy in from entire organization.

Apache **Hadoop** software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

From reading PDF –Data Mining

Data mining – Identify key questions considering cost trade off of exercise for higher level of accuracy

- need right data – may already be available but may have to produce it (Ex: surveys)

- have to preprocess data and deal with missing data in a formal way. If missing data is systematic then need to determine potential impacts on results

- Transforming data – determine appropriate format for data storage and scale down attributes including transforming variables

- Storing data in format conducive to mining and on secure mediums

- Mining data after data is processed, transformed, and stored and is subject to data analysis algorithms and machine learning. good starting point is visualization.

- Evaluate the results. ***in-sample forecast:*** testing predictive capabilities of models on data to see if algorithms in data reproduction were effective. Results are shared with key stakeholders.

**Module 4: Deep Learning and Machine Learning**

Big Data – huge data sets think 5V’s

Data mining – analysis to see patters (see above)

Machine learning – uses algorithms to analyze data and makes intelligent decisions without explicit programming – allows to solve problems and make accurate predictions on their own.

Deep learning – subset of machine learning – uses layered neural networks to sim human decision making. Can label, categorize info and ident patterns. Enables AI to continuously learn on the job.

Neural network – computing units called neurons that take data and learn to make decisions.

Artificial Intelligence vs Data Science:

DS – process of extracting knowledge/insights from data with math, stats, visualization, uses machine and deep learning.

AI - is everything that allows computers to learn how to solve problems and make intelligent decisions.

Deep learning – need basic linear algebra understanding.

Application of machine learning

– Predictive analytics: decision trees, Bayesian analysis, Naïve Bayes

- Recommendations – like Netflix – you might like this… also used in fintech (financial technology) – recommendations for other investments, or in areas like fraud detection

**Reading – Regression**

**-** can quantify magnitude of relationships.

**Module 5 – data Science in Business**

Health care – uses modeling, mining, stats, and predictive analytics

Natural disaster predicter

Companies – collect, archive, measure data

Application of DS – recommendation engines (amazon suggestion stuff, netflix)

Reading – the Final Deliverable

Purpose of analytics – deliver insight to those who who need it for informed decision making, summarizes findings, used to build narrative. Equivalent to research papers in academia

Deliverable needs narrative!

**Module 6 – how to become DS**

- Highest end DS’s are PHD’s with comp science background

Basic programming, algebra, calculus, stats, database (relational)

Recruiting –

Reading – The Report Structure

If brief then generally summary of findings (ex: commentaries on current trends)

but if long then build with methodology and sources… etc. (ex: critical reviews and analyzing/commentary

\*\*\* ALWAYS include a cover page! (title, names, publisher, data of pub)

If short – still recommend exec summary or abstract

Intro – including relevant research and related literature review

Methodology – data sources and research methods

Results section – empirical findings

discussion – craft main arguments and highlight findings

conclusion – general specific findings and mention potential beneficial research

Checklist – HAVE YOU????

from start write what reader gains by reading

clearly stated aim of work?

Explain signif of contribution

given sufficient background

addressed question of usefulness

identified future develeopments

appropriately structure paper

**COURSE 2 – DS Tools**

**Mod 1 –**

Which Languages should I learn? no right answer. Depends on needs, problems and for who,

can also depend on age of application and role in company

python, R, and Sequel recommended by IBM

Intro to python – most popular dor DS – 75% of ds say they use on regular basis. 75% of job descriptions included python: open sources (based on business), good for starters and experienced programmers, good overall general language

good for data science, AI, machine learning

Facebook, nasa, google, amazon

- large standard database (panda, Numpy), general overall language,

- AI – ex: Pytorch

- has huge global inclusive community

Intro to ‘R’: (learn up to 3 languages to increase salary)

- free software ( based on values)

- public collaboration, private use or business use

- statisticians, data miners

- easier to translate from math to code

- intermidate language

- IBM, google, facebook, uber

- largest repository of statistical knowledge

- integrates well with other languages

- big global communities

Intro to SQL: must in DS and data engineering

- data accessed directly with no prior need to copy.. accelerates workflow executions

- Structured Query Language

- non procedural language and scope limited to ‘querying’ and ‘managing’ data

- powerful and older that python and R by about 20 years.

- developed by ibm in 1974. handles structured data (data incorporating realtions btw entities and variables.

- managing data in relational databases. (formed by collections of two dimensional tables

- The SQL language is subdivided into several language elements, including clauses, expressions,

predicates, queries, and statements.

- Knowledge can be applied to other databases

- MYsql, IMBdb2,

sql writing depends on relational database in use

Other languages: Scala, java, c++ , juli

jave - general-purpose object oriented programming language. fast and scalable, run on java vritual machine.

- javamL – machine learning librabry made from java

- java also built apache Hadoop, It manages data processing and storage for big data applications running in clustered systems.

Scala runs on JVM and works with java

For DS – apache spark is common

C++ - general prog language – improvse process speed.

- Tensorflow – popular deep learning built with C++ but runs on python interface so don’t need to know C++

- javascript – tensorflow.js for DS

**Categories of DS tools**

- Data mgmt – processing/retrieving data,

- Data integration/transformation or ETL (Extract, Transport, Load) from remote db’s

- data visualization – final deliverable

- model buiding, deep learning machine

- model deployment – makes deep machine available to 3rd party

- model monitoring – quality checks

- data asset mgmt – brings same versioning to data

- development enviros – IDE’s

- execution enviros’ – preprocessing, model training

- Development environments, commonly known as Integrated Development Environments, or “IDEs”,

are tools that help the data scientist to implement, execute, test, and deploy their

**Open source tools for DS**

- data mgmt – common are relational dbs such as MySQL and PostgreSQL; NoSQL databases such as MongoDB Apache CouchDB, and Apache Cassandra; and file-based tools such as the Hadoop File System or Cloud File systems like Ceph. Elastiserch for storing text and indexing for fast retrieval.

- ELT is now ETL and new term emerged called data refining and cleansing

- Data integration and transformation tools

Apache AirFlow, originally created by AirBNB; KubeFlow, which enables you to execute data science pipelines on top of Kubernetes; Apache Kafka, which originated from LinkedIn; Apache Nifi, which delivers a very nice visual editor; Apache SparkSQL (which enables you to use ANSI SQL and scales up to compute clusters of 1000s of nodes), and NodeRED, which also provides a visual editor. NodeRED consumes so little in resources that it even runs on small devices like a Raspberry Pi.

**-**Data visulation tools: Hue, which can create visualizations from SQL queries. Kibana, a data exploration

and visualization web application, is limited to Elasticsearch the data provider). Finally, Apache

Superset is a data exploration and visualization web application.

-Model deployment – (deployment of machine learning models) - Apache predictionIO, Seldon is an interesting product since it supports nearly every framework, including TensorFlow, Apache SparkML, R, and scikit-learn

- Model Monitoring tools - ModelDB is a machine model metadatabase where information about the models are stored and can be queried. A generic, multi-purpose tool called Prometheus is also widely used for machine learning model monitoring, although it’s not specifically made for this purpose.

- IBM AI Fairness 360 open source toolkit does exactly this. It detects and mitigates against bias in machine learning models.

- IBM Adversarial Robustness 360 Toolbox can be used to detect vulnerability to adversarial attacks

- IBM AI Explainability 360 Toolkit makes the machine learning process more understandable by finding similar examples within a dataset

- Code asset mgmt (aka version control) – github, gitlab, bitbucket

- data asset mgmt or data lineage: data has to be versioned and annotated with metadata. Apache Atlas, kylo,

**Open Source Development tools**

Development enviro – Jupyter, supports over 100 programming languages or kernels.

Jupyter Notebooks is the ability to unify documentation, code, output from the code, shell commands, and visualizations into a single document.

- Apache Zeppelin, like jupyter notebooks but with integrated plotting capability

- Rstudio 2011 – runs R and R librabries, unifies programming, execution, debugging, remote data access, data exploration, and visualization into a single tool.

- Spyder tries to mimic the behavior of RStudio to bring its functionality to the Python world.

Cluster Execution enviros – Apache Spark, - key is lenear scalabitliy (if u double servers in cluster than you double its performance. pache Spark is a batch data processing engine, capable of processing huge amounts of data file by file.

- apache flink: is a stream processing image, with its main focus on processing real-time data streams.

- ray – Large scale deep learning model training

Fully integrated and visual (no programming necessary): Knime

orange

**Commercial Tools for DS**

- data management:

- most of an enterprise’s relevant data is stored in an Oracle Database, Microsoft SQL Server, or IBM Db2.

- generally referring to ETL tools

- Informatica Powercenter and IBM InfoSphere DataStage are the leaders, followed by products from SAP, Oracle, SAS, Talend, and Microsoft.

- these tools support design and deployment of ETL data-processing pipelines through a

graphical interface.

Data visualization:

- main purpose is to produces visually appealing reports and dashboards – Power BI, Tableau, IBM Cognos

Machine learning model: SPSS modeler, SAS Entreprise Miner

Model Montoring – relatively new and no commercial products yet

Asset mgmt – same as above. open source is best

Data asset mgmt to version and annotate – Informatica, IBM

Watson Studio is a fully integrated development environment for data scientists. Watson studio together with Watson Open Scale, is a fully integrated tool covering the full data science life cycle

**Cloud Based Tools for Data Science**

**-** have multiple tasks in single tool

Fully integrated visual tool

- Watson studio, together with Watson Open scale covers the complete development lifecycle for all data science, machine learning and AI tasks.

MS Azure, and H2O driverless AI

SAS = Software as a service. cloud provider operates tool for you in cloud

Amazon Web Services Dynamo DB, a no SQL database that allows storage and retrieval of data in a key value or a document store format.

Jason – most prominent document data structure

Two widely used commercial data integration tools are Informatica Cloud data integration

and IBM's data refinery data refinery

enables transformation of large amounts of raw data into consumable quality information in a spreadsheet like user interface.

Machine Learning – Watson Machine Learning

Amazon sage maker model monitor is an example of a cloud tool that continuously monitors

deployed machine learning and deep learning models.

**Libraries for DS:**

- Libraries are a collection of functions and methods that enable you to perform a wide variety of actions without writing the code yourself.

- Libraries usually contain built-in modules providing different functionalities that you can use directly; these are sometimes called “frameworks.”

- Pandas offers data structures and tools for effective data cleaning, manipulation, and analysis.

- Data vis libraries - Matplotlib package excellent for graphs and plots, Seaborn

- Machine learning libs – Scikit for statistical modelling, regression, classification

- Deep learning libs – Keras, tensorflow for production and experimentation, pytorch for experimentation

- Apache Sparks processes data through compute clusters (process in parallel using many computers simultaneously. Can use python, R, or SQL

- Data enginerring libs – Scala

- Deep learning – BigDL,

R and python are the standard for open source DS

**Application Programming Interfaces (API)**

- API’s allow two pieces of software to talk

- API only refers to interface (or library you see), Library refers to the whole thing

- REST (Representational State Transfer) APIs are another popular type of API. They enable you to communicate using the internet, taking advantage of storage, greater data access, artificial intelligence algorithms, and many other resources.

Your program is called the client. API communicates with web, sets of rules govern comms

Common API terms – Code = client

Web service = Resource

HTTP = transmits over the internet (usually a JSON file to instruct service to perform. Info Transmitted back to client

Ex of REST API – Watson speech to text API

API converts speech to text. u send copy of audio to API. API then sends transcript of what individ says.

**Data Sets - Powering Data Science**

- A structured set of data (various types of info) – structured as tabular data with rows and columns

Ex: CVS – comma separated values (a delimited text file where each line reps a row and data values

Hierarchichal or network data structures are used to rep relations btw data. Hierarchical is tree-like structure, network data is stored as a graph

- can include raw data files, EX: MNIST (used to train image processing systems)

- usualy contain sensitive data

- becoming more popular for GOv bodies to release open data, used publicly and commercially to create new apps, or for research, Open Knowledge Foundation’s datacatalogs.org website. Kaggle, google.

- can be defined by licencing terms

for distribution - Linux Foundation created the Community Data License Agreement, or CDLA. –

- CDLA sharing licence grants permission to modify and use but under same licence terms of original data

- CDLA permissive licence – use and modify but don’t have to share changes made to data

-

**Data Asset Exchange**

- DAX created by IBM, - trusted source to find quality open source data, licences and terms. Try to make available under CDLA,

- includes data sets from IBM research

- has notebook tutorials

- Most data sets on DAX are complemented by one or more Jupyter Notebooks that you can

use to perform data cleaning, pre-processing, and exploratory analysis.

**Machine Learning Models:**

- uses algorithms or models to ident patterns

**MODULE 2: OPEN SOURCE TOOLS**

**Overview of Github:**

- popular fro DS and developers - for performing version control of source code files and projects

and collaborating with others.

- version control to keep track of changes

- Git is free and under GNU (general public licence

- mainly for code but can do documents, images, and other file types

- github is most popular for git repositories, also gitlab, bitbucket, and beanstalk

- SSH protocol is a method for secure remote login from one computer to another.

- try.github.io to download the cheat sheets and run through the tutorials.

-

**Github part 1:**