Notes about IBM Data Analyst Professional Certificate

Douglas Cardoso

March 2021

Parte I

Introduction to Data Analytics

1 Overview of Data Repositories

Databases: collection of data or information. DBMS: Data Base Management System é o conjunto que cria e mantem os databases. Permite você armazenar, modificar e extrair dados dos databases. Há diferentes tipos, os principais são: relational e non relation. Relational também é conhecido como RDBMS, baseado em dados tabulares, usa-se SQL para dar query nos large data dos databases. Os non-relational (NoSQL), permite armazenar dados mais modernos como audios, midias sociais e outros.

Temos também as Data Warehouse, que é um repositório de informações de diferentes fontes e consolida com o ETL (extract, transform and load). ETL ajuda a extrair os dados das diferentes databases, de forma clean e que possa ser utilizado.

1.1 RDBMS

Relational database is a collection of data organized into a table structure, with columns(records) and rows (attributes). Permite que você faça querys em várias tabelas. Baseiam-se em organizational principles of flat files, such a spreadsheet. Use SQL for querying data. Optimize storage for large values of data, has a unique set of rows and columns, relationships can be defined between tables, and others advantages.

Podem ser open-soure with internal support or with commercial support and commercial closed-source. Tem várias exemplos de RDBMS. Umas das vantagens são: minimize data redundancy, join tables, flexibility, ease backup, ACID Compliance (confiabilidade).

1.2 NoSQL

Not Only SQL is a non-relation database desing, feitos para dados modernos de alta flexibilidade, com structure diferentes. Há quatro tipos comuns de NoSQL databases: Key-value store, Document Based, Column Based and Graph Based. Suas vantagens são: ability to handle large volumes of structure, semi and unstruc data; desing simples; datacenters que permite conciliação com as clouds e outros.

Key-value store são armazenados como uma coleção de pares valores chaves, sendo a chave um atributo e identificador único do dado. As keys podem ser qualquer coisa. Os conhecidos são: Redis, Memcached and DynamoDB.

Document Based são database que armazenam cada registro e seus dados associados em um único documento Permite indexação flexível, consultas ad hoc e análises das collections of documents. Os mais usados são: MongoDB, DocumentDB, CouchDB, Cloudant.

Column Based são databases que armazenam dados em células agrupadas por columas em vez de linhas, a logical group of columns is call a column family, acessar os dados se tornam super rápido. Os mais populares são: Cassandra and APACHE HBase.

Graph base são usam modelos gráficos para armazenar os dados, procuram analisar e encontrar connections between differents peaces of data. Os mais populares são: neo4j and CosmosDB.

1.3 Data Marts, Data Lakes, ETL, and Data Pipelines

Data Warehouse: funciona como um armazenamento multiuso, os dados já chegam estruturados para seu propósito, já estão pronto para análise. É como uma fonte verdadeira, storage current and historic datas for analysis.

Data marts: subsection of warehouse bulding for a specific business function, propósito. Oferece recursos analíticos, emissão de relatórios.

Data lake: armazena grande volume de dados de todos os tipos, recebe um identifier exclusivo e é tagged with metatags for further use. É usado pra gerar dados contínuos sem propósito específico, retendo todos os dados sem excluir nada. Análise preditiva e avançada.

ETL: Extract, Transform and Load. Clean, standard and transform datas for report and analysis. Para extraction de data em intervalos específicos vindo da origem, usa-se duas ferramentas principais: BLEND and Stitch. Para extraction em tempo real incluem APACHE Samza, STORM, and kafta. Transform deixa o dado limpo para a análise, sem duplicação ou NaN. Load é etapa de quando os dados processados são transportados para um sistema de destino ou data repository. Load verification é importante para consultar se está tudo certo com os dados processados.

Data pipeline: (engloba) encompasses the entire journey of moving data from one system to another, including the ETL process. O destino normalmente é uma data lake.

1.4 Foundations of 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 drive real-time business insights that relate to consumers, risk, profit, performance, productivity management, and enhanced shareholder value."
>
> (Ernst & Young)

The V's of Big Data: velocity, volume, variety, veracity and value. Velocidade está se gerando rapidamente, em um volume gigantes, de uma variedade de tipos de dados, structured and unstructured, vindos de diferentes fontes, incluindo elementos de qualidade de veracidade do dado (a informação é verdade ou falsa?), que trazem valores de diversos tipos para diversas pessoas. Todos os diversos dispositivos capturam e armazenam approximately 2.5 quintillion bytes every day. Ferramentas como Apache Spark e Hadoop ajudam a lidar com esse volume gigantesco de dados.

1.4.1 Big Data Processing Tools

There is three open source technology that lead with Big Data, as Apache Hadoop, Apache Hive and Apach Spark. Hadoop is a collection of tools that provide distributed storage and processing big data. Hive is a data warehouse for data query and analysis. Spark is a distributed analytics framework for complex real-time data analytics.

Hadoop trabalha com clusters de nós de vários computadores, fornecendo self-service in real time, and etc. Ele usa o sistema de armazenamento HDFS, escalável e confiável, com multiple nods in multiple computers, com calculos feito em pararelos. Além disso, by default, Hadoop replica as peças menores em dois servidores adicionais, garantindo segurança caso um falhe. Parece muito com o sistema de clouds da AWS.

Hive is a open-source para leitura, gravação e management of large datasets files. É mais adequado para ETL e inclui tools que permitem o acesso mais fácil através o SQL.

Spark é projetado para extract and process large volume of data, para várias coisas, como interactive analytics, streams processing, machine learning, data integration and ETL. Tem interfaces para a maioria das linguagens de programação.

2 Identifying Data for Analysis

O primeiro passo é definir a informação/dados que você quer coletar para resolver/entender o seu problema, definindo também as fontes. Após isso, você precisa definir um plano para a coleção dos dados, definindo, por exemplo, o timeframe, quantos dados

serão necessários, além de poder definir os riscos e dependecies. O terceiro passo é determinar os métodos que você usará para coletar os dados, os sources, os types, timeframe and volume.

É preciso certificar-se de que os dados são de qualidade, free of erros, accurate, complete, relevant and acessible. Além disso, os dados precisam ser seguros e tomar cuidado com data privacy.

2.1 Data Sources

Data sources pode ser internal or external, e também primary (data pegos diretamente por você na fonte), secondary (information retrieved from existing sources) and third-party (dados que tu comprou de alguma empresa que vende databases agregadas de vários outros sources).

2.2 Data Wrangling

O processo de data wrangling consiste em quatro etapas: discovery, transformation, validation and publishing. O primeiro passo tem o objetivo de identificar a melhor forma de limpar e transformar os dados que você tem na mão. A fase de *transformation* se divide em quatro etapas principais: **structuring** (muda os formatos e schemas dos dados, usando join and unions, combine data for two or more databases sources, unions combine rows and joins combine columns), **normalization and denormalization** (cleaning unsed data, reducing redudancy and inconsistency), **cleaning data** (fixing irregularities, leading with inaccuracies, missing data, incomplete data, biases in data, null values and outliers), and **enriching data** (adding data points). Após isso temos a fase de *validation*, que checa a qualidade dos passos anteriores feitos em *transformation*. Por fim, temos o *publishing*, que é entregar os dados depois do weangling para as necessidades do projeto em questão.

2.3 Tools for data wrangling

Sobre os principais softwares usados em data wrangling, temos como principais:

- Excel Power Query/Spreadsheets
- OpenRefine
- Google DataPrep
- Watson Studio Refinery
- Trifacta Wrangler
- Python and R

2.4 Data Cleaning

Um típico processo de data cleaning inclui inspection, cleaning and verification. A inspection é para detectar qualquer issue and erros em seus dados, usando data visualization using statistical method and data profile (structure, content and inter-relationships of your datas).

A parte de cleaning depende do que você encontrou, caso tenha sido missing value, você pode filtrar ou encontrar os source missing information, assim como usar imputation. Você pode ter duplication data, que precisa, nesse caso, be removed. Assim como, você pode encontrar irrelevant data. Inclui aqui data type transformation, standardizing and lead with syntax erros. Outliers precisam ser examinados, pois podem estar corretos ou incorretos, decidindo assim pela inclusão ou exclusão desse data point em sua database.

Já a validation, você inspecta seus dados procurando identificar a accuracy and precision achieved as result of the data cleaning. É preciso lembrar que, todas as mudanças feitas precisam ser documentadas.

3 Analyzing and Mining Data

Primeiramente, o que é statistics? É um ramo da matemática que lida com coleta, análise, interpretação e apresentation de dados. A análise estatística é a aplicação da estatística em uma amostra de dados com o objetivo de decifrar ou entender o que esses dados representam. Uma amostra (sample) é uma seleção representativa extraída de uma população total, onde a população é um grupo de pessoas ou coisas que podem ser identificadas por uma ou mais características.

Dentro da estatística, temos a descritiva, que permitem que façamos análises básicas sobre nossa amostra, com uso de algumas métricas, como central tendency (mean, median and mode), skewness (measure of whether the distribution of values is symmetrical arount a central value or skewed left or right) and dispersion (variance, std and range). Outro tipo de statistical analysis is the inferential statistics, que usa os dados da amostra pra fazer previsões about data of population. Na prática, você desenha generalizações na amostra que são aplicáveis a população como um todo. Incluem ai hypothesis testing (study of effectiveness), confidence intervals (incorporate the uncertainty and sample error) and regression analysis (incorporate hypothesis tests that help determine whether the relationships observed in the sample data actually exist in the population rather than just the sample). Dentre os softwares para realizar statistical analysis estão SAS (Statistical Analysis System), SPSS (Statistical Package for the Social Sciences) ou Stat Soft. Statistical, combined with Data Mining, togheter help in better decision-making.

- Providing measures and methodologies necessary for data mining

- Identifying patterns that help identify difference between random noise and significant findings.

3.1 What is Data Mining?

Data mining is the process of extracting knowledge from data. Involve recognition technologies, statistical analysis, and mathematical techniques.The goal is identify correlation, find patterns and variations, understand trends and predict probabilities. Pattern recognition is the discovery of regularities, or commonalities, in data. A trend is the general tendency of a set of data to change over time.

Dentre as técnicas mais comuns de data mining estão:

- Classification - Classifying attributes into target categories
- Clusteiring - Involves grouping data into clusters so they can be treated as groups
- Anomal or Outlier Detection
- Association Rule Mining - Establishing a relationship between two data events
- Sequential Patterns - tracing a series of events that take place in a sequence
- Affinity Grouping - Discovering co-occurrence in relationships
- Decisions Trees - Bulding classification models
- Regression - Help identify the nature of relationship between two variables, causal or correlational

3.2 Tools for Data Mining

Some of the commonly used software and tools for data mining.

- Spreadsheets (Data Mining Client, XLMiner, KnowledgeMiner)
- R (tm, twitteR)
- Python (pandas, NumPy, jupyter)
- IBM SPSS Statiscs (
- IBM Watson Studio
- SAS

3.3 Overview of Storytelling

É necessário ter a habilidade de contar o que os gráficos e tabelas estão dizendo. Para isso há alguns passos importantes de se entender:

- Who is my audience?
- What is important to them?
- What will help them trust me? Falar a mesma língua e mostrar que entende do problema do qual está falando

3.4 Data Visualization

É importante comunicar seus dados de forma visual, com graphs, charts, and maps. Há vários tipos de data visualization, e você precisa escolher a que melhor comunica os dados que você tem. Tente responder a pergunta: "What is the question I'm trying to answer?". Alguns tipos:

- Bar charts: are great to comparing related data sets of parts of a whole.
- Columns charts: compare values side-by-side
- Pie charts: show the breakdown of an entity into its sub-parts and the proportion of the sub-parts in relation to one another.
- Line charts: show how a data value is changing in relation to a continuous values
- Dashboards: organize and display reports and visualizations coming from multiple data sources