# Big Data
## (Analytics)

Instructor: Thanh Binh Nguyen

September 1st, 2019



S³Lab
*Smart Software System Laboratory*


“Big data is at the foundation of all the megatrends that are happening today, from social to mobile to cloud to gaming.”
– Chris Lynch, Vertica Systems

# What is Data Analytics

- **Data on its own is useless unless you can make sense of it!**
- Data or information is in raw format. The increase in size of the data has lead to a rise in need for carrying out inspection, data cleaning, transformation as well as data modeling to gain insights from the data in order to derive conclusions for better decision making process. This process is known as **data analysis**.

# What is Data Analytics

- **Data Mining** is a popular type of data analysis technique to carry out data modeling as well as knowledge discovery that is geared towards predictive purposes.
- **Business Intelligence** operations provide various data analysis capabilities that rely on data aggregation as well as focus on the domain expertise of businesses. In Statistical applications, business analytics can be divided into **Exploratory Data Analysis** (**EDA**) and **Confirmatory Data Analysis** (**CDA**).

# What is Data Analytics

*Reporting vs Analysis*

- A **report** will show the user what had happened in the past to avoid inferences and help to get a feel for the data while **analysis** provides answers to any question or issue. An analysis process takes any steps needed to get the answers to those questions.
- **Reporting** just provides the data that is asked for while **analysis** provides the information or the answer that is needed actually.

# What is Data Analytics

*Reporting vs Analysis*

- We perform the **reporting** in a standardized way, but we can customize the **analysis**. There are fixed standard formats for **reporting** while we perform the **analysis** as per the requirement; we customize it as needed.
- We can perform **reporting** using a tool and it generally does not involve any person in the analysis. Whereas, a person is there for doing **analysis** and leading the complete **analysis** process.

# What is Data Analytics

*Reporting vs Analysis*

- **Reporting** is inflexible while **analysis** is flexible. **Reporting** provides no or limited context about what's happening in the data and hence is inflexible while **analysis** emphasizes data points that are significant, unique, or special, and it explains why they are important to the business.

# Applications of Data Analytics

# The case for Business Analytics / BI

**BUSINESS NEED**

- The Business environment today is more complex than ever before.
- Businesses are expected to be diligently responsive to the increasing demands of customers, various stakeholders and even regulators.

**SOLUTION**

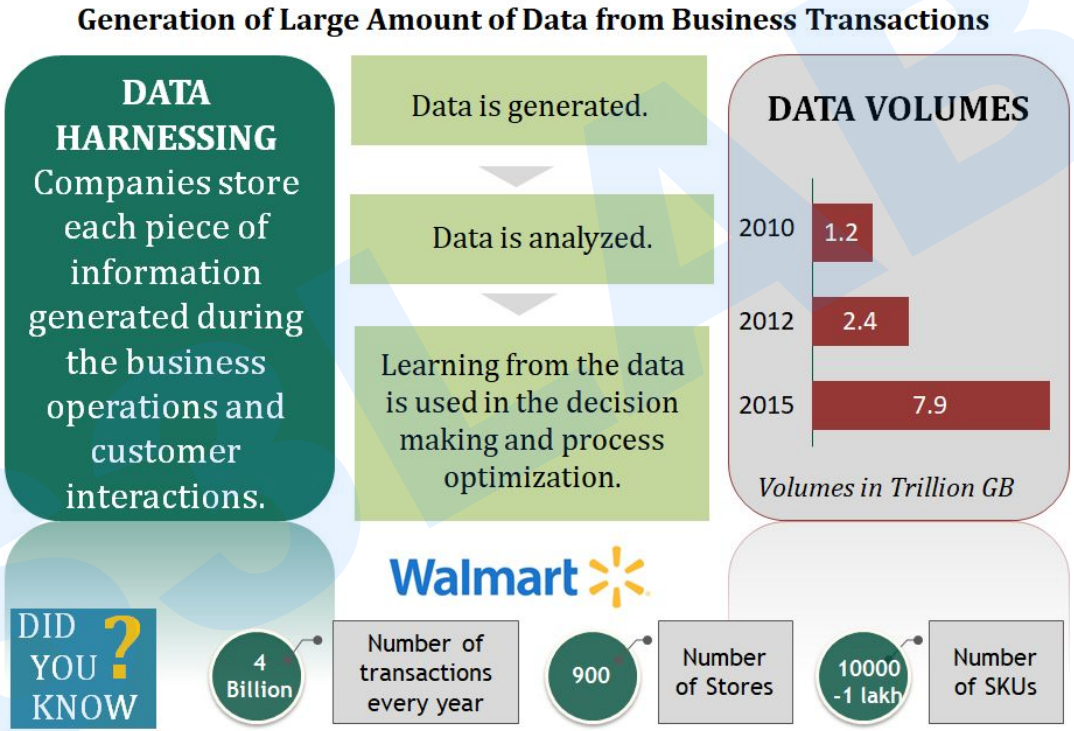
- Organizations have been turning to the use of analytics.
- More than 83% of Global CIOs surveyed by IBM in 2010 singled out Business Intelligence and Analytics as one of their visionary plans for enhancing competitiveness.

**GOAL**

In most cases the primary objective of an organization that seeks to turn to analytics is:

- Revenue/Profit growth
- Optimize expenditure

# Grow Needs for Analytics



# Why Big Data analytics?


### Why is Big Data Analytics important?

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

# Data Mining

- Data mining also called data or **knowledge discovery** means analyzing data from different perspectives and **summarizing** it into **useful information** – information that we can use to make important decisions. It is the technique of exploring, analyzing, and detecting **patterns** in large amounts of data. The goal of data mining is either data classification or data prediction. In classification, we sort the data into groups while in prediction, we predict the value of a continuous variable.

# Data Mining

*Examples*

- Classification of Trees
- Logistic Regression
- Neural Networks
- Clustering Techniques like the K-nearest Neighbours
- Anomaly Detection

# Types of Analytics

# Types of Analytics

- Descriptive Analysis
  - With the help of descriptive analysis, we analyze and describe the features of a data. It deals with the summarization of information. Descriptive analysis, when coupled with visual analysis provides us with a comprehensive structure of data.
  - In the descriptive analysis, we deal with the past data to draw conclusions and present our data in the form of dashboards. In businesses, descriptive analysis is used for determining the Key Performance Indicator or KPI to evaluate the performance of the business.

# Types of Analytics

- Predictive Analysis
  - Based on the analysis of the historical data, we are able to forecast the future. It makes use of descriptive analysis to generate predictions about the future. With the help of technological advancements and machine learning, we are able to obtain predictive insights about the future.
  - Predictive analytics is a complex field that requires a large amount of data, skilled implementation of predictive models and its tuning to obtain accurate predictions. This requires a skilled workforce that is well versed in machine learning to develop effective models.

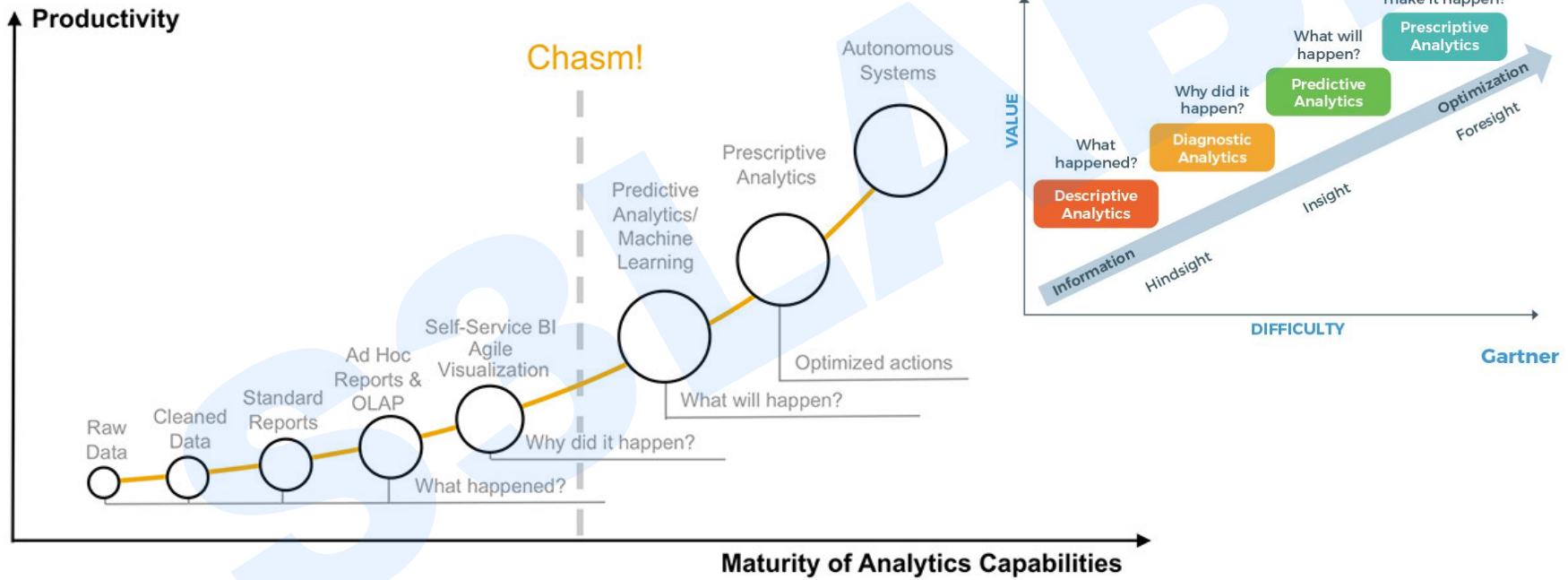
# Types of Analytics

- Diagnostic Analysis
  - At times, businesses are required to think critically about the nature of data and understand the descriptive analysis in depth. In order to find issues in the data, we need to find anomalous patterns that might contribute towards the poor performance of our model.
  - With diagnostic analysis, you are able to diagnose various problems that are exhibited through your data. Businesses use this technique to reduce their losses and optimize their performances. Some of the examples where businesses use diagnostic analysis are:
    - Businesses implement diagnostic analysis to reduce latency in logistics and optimize their production process.
    - With the help of diagnostic analysis in the sales domain, one can update the marketing strategies which would otherwise attenuate the total revenue.

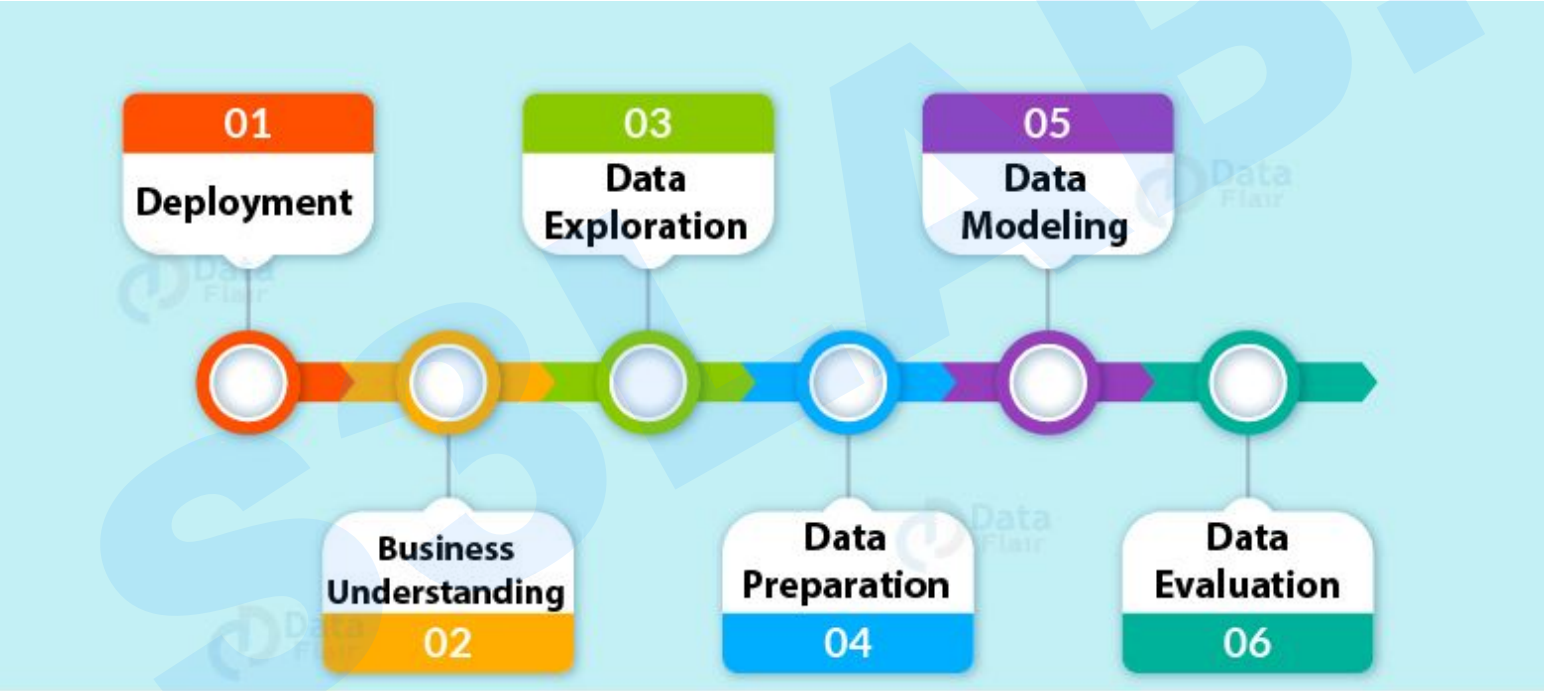
# Types of Analytics

- Prescriptive Analysis
  - Prescriptive analysis combines insights from all of the above analytical techniques. It is referred to as the final frontier of data analytics. Prescriptive analytics allows companies to make decisions based on them. It makes heavy usage of Artificial Intelligence in order to facilitate companies into making careful business decisions.
  - Major industrial players like Facebook, Netflix, Amazon, and Google are using prescriptive analytics to make key business decisions. Furthermore, financial institutions are gradually leveraging the power of this technique to increase their revenue.

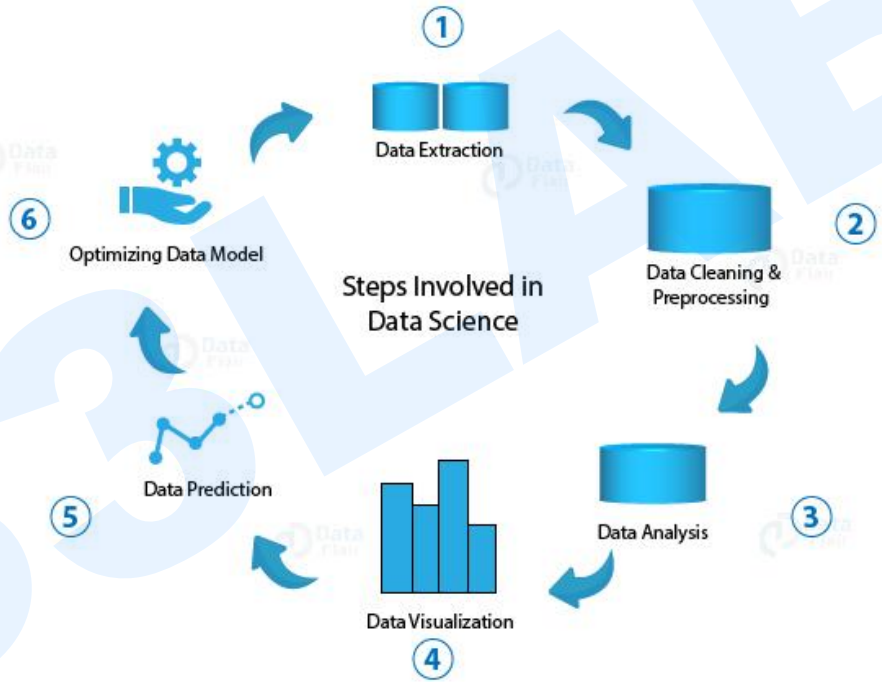
# Types of Analytics



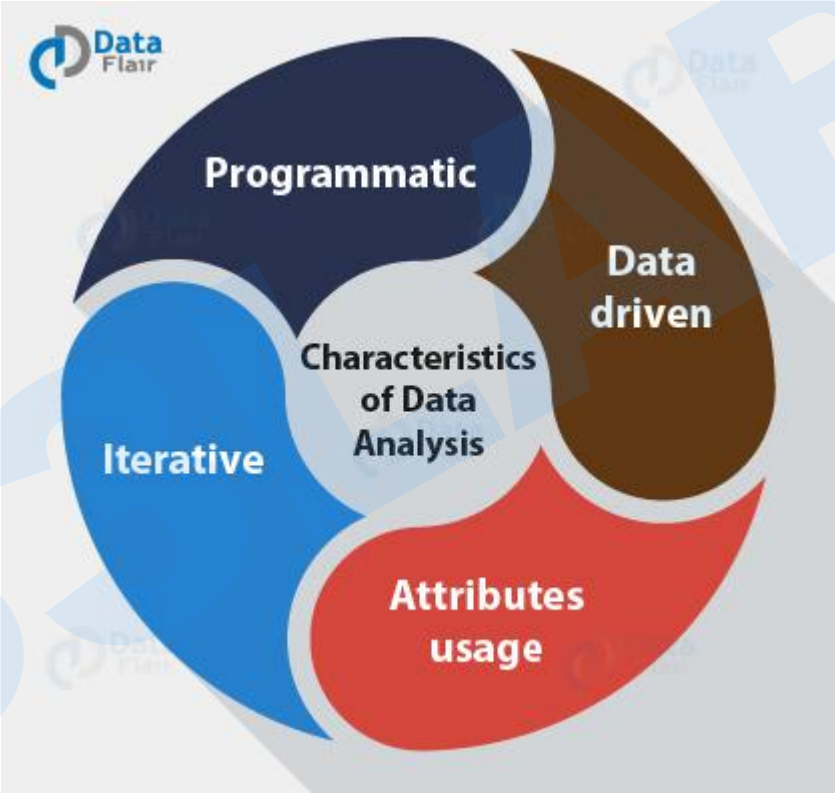
# Process of Data Analysis



# Process of Data Analysis



# Characteristics of Data Analysis



# Data Scientist

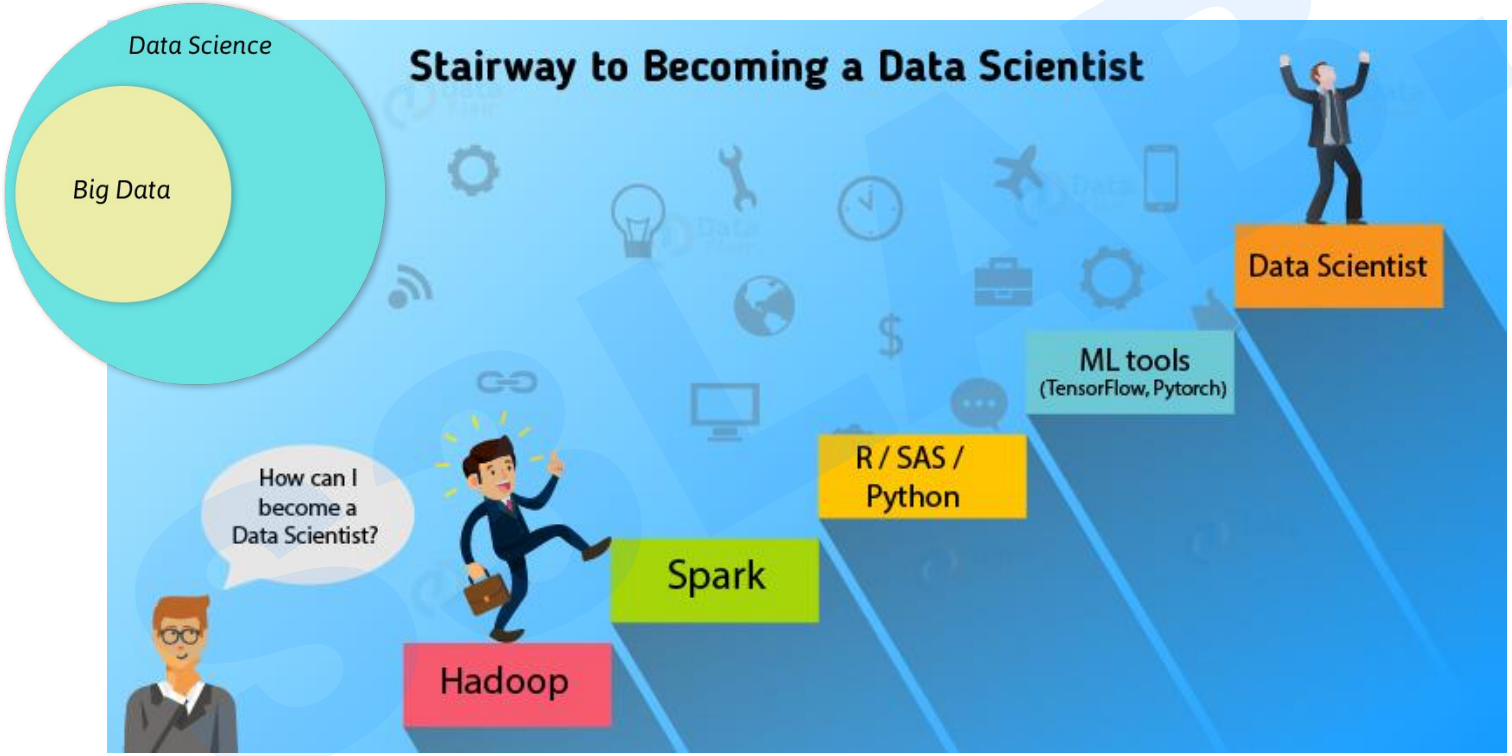

A Business analyst is not able to discover insights from huge sets of data of different domains.

Data scientists can work in co-ordination with different verticals of an organization and find useful patterns/insights for a company to make tangible business decisions.

**15,000%**

INCREASE IN JOB POSTINGS FOR DATA SCIENTISTS IN THE US BETWEEN 2011-12

# Data Scientist



# Know your Tools

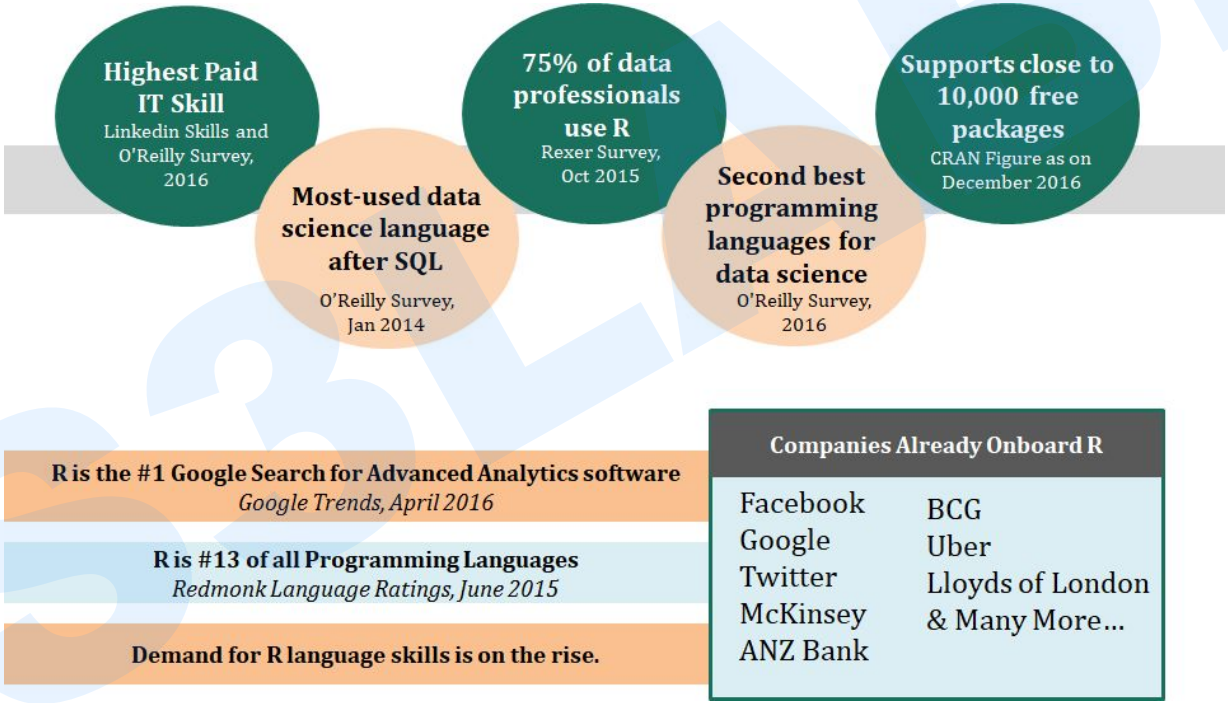

The program is developed keeping in mind the needs of an evolving Analytics industry that requires individuals to be "job-ready" from Day 1.

# Know your Tools

## SAS

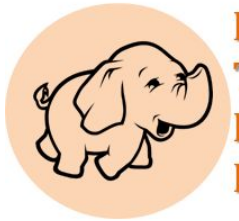
# Know your Tools

*R*



# Know your Tools

## *Hadoop*



**Hadoop is Transforming Businesses Across Industries**


*Organizations use Hadoop to manage their data today (up from 1 out of 10 in 2012)*


**BIG DATA STORING AND FASTER PROCESSING**

Hadoop is an open source software framework created in 2005 that keeps and processes big data in a distributed manner on large collection of hardware.

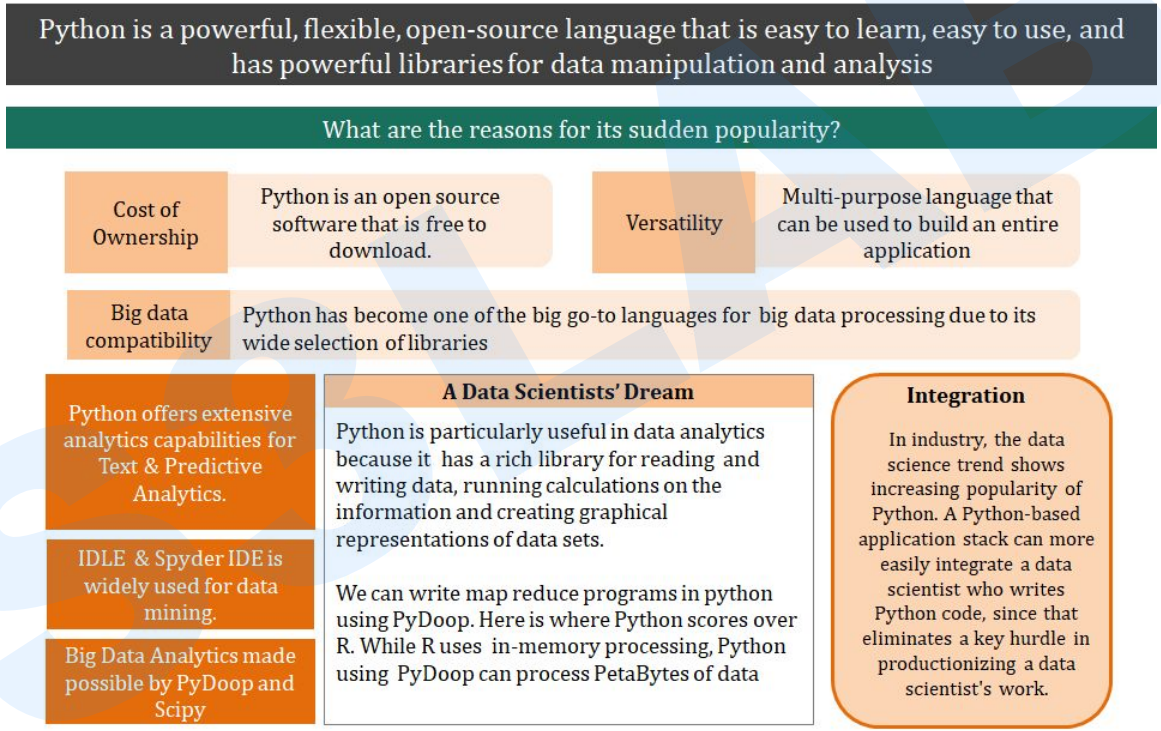
**BUSINESS SOLUTIONS ACROSS DOMAINS AND INDUSTRIES:**

Low cost solution with a high fault tolerance to access and create value from data.

*"The growing use of Apache Hadoop, increasing data warehouse volume sizes and the accumulation of legacy systems in organizations are fostering structured data growth. These factors are leading enterprises to understand how to reuse, repurpose and gain critical insight from this data."* Gartner

# Know your Tools

## Python



# Know your Tools

## Python

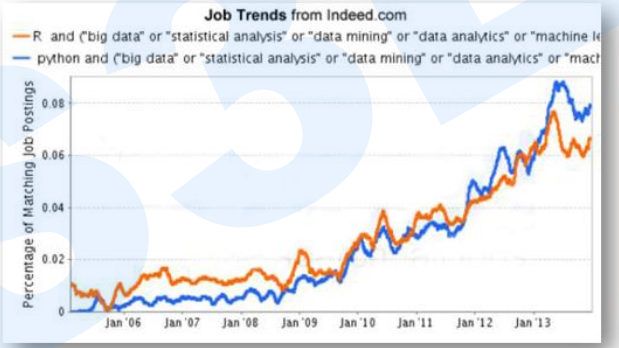
**Official language of Google**

**Among top in-demand data science skills**
KDNuggets, Dec2014

**46% of job ads mention Python (after SQL)**
KDNuggets Dec 2014

**Ranked #1 of all programming languages**
Codeeval rankings, Feb 2015

**2nd most popular data science language**
KDNuggets 2013




**Companies Already Onboard Python**

| | |
|---|---|
| Google | IBM |
| Yahoo | National Weather |
| Quora | Service |
| Nokia | & Many More… |
| ABN | |
| AMRO Bank | |

# Know your Tools

*Tableau - what is data visualization*

Data visualization is the presentation of data in a pictorial or graphical format. For centuries, people have depended on visual representations such as charts and maps to understand information more easily and quickly.

# Know your Tools

*Tableau - what is needed in a good visualization tool?*

- Every data visualization tool is expected to transform unfathomable data into eye-pleasing charts and graphs and let them convey the hidden message in the data to the business users and analysts.
- Rather, a tool must also have the capability to analyze, process and present the data in a digestible manner.
- The tool must be efficient in creating understandable data reports and dashboards by assimilating and aggregating massive and complex data from different sources.

# Know your Tools

*Tableau - what is needed in a good visualization tool?*

- The tool must be so designed that every user, regardless of their background and skill set can learn to use it like any other software or application used daily.
- A tool must focus on being an interactive self-service analytics, development, and visualization tool. This is unlike the traditional enterprise-wide, IT-developed software which can only be used by a small group of technical users. A tool should give entry-level users the capability to use it efficiently and gain insights into big data.

# Know your Tools

*Tableau - what is needed in a good visualization tool?*

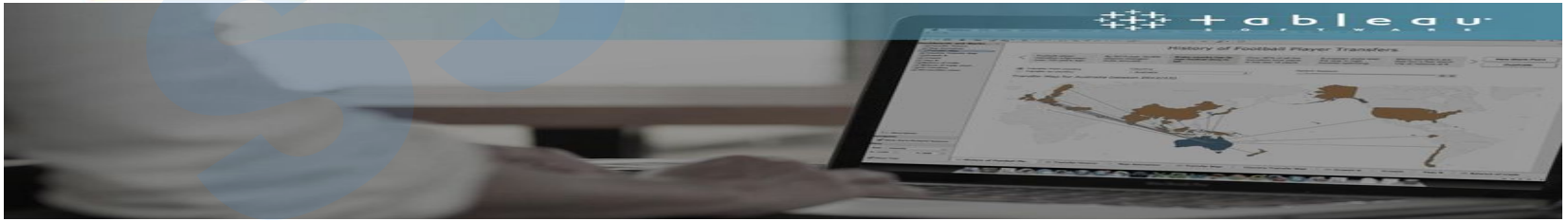
- Filtering, processing, and slicing-dicing of big data should be easy.
- A tool must provide enough features for **collaboration** and insight **sharing** within and outside the organization.
- The visualizations must be plenty and intuitive for any technical or non-technical user to understand it and draw meaningful insights from it.

# Know your Tools

*Tableau*

**Tableau is a powerful, flexible Data Visualization tool that is easy to learn, easy to use, and has powerful libraries for data visualization and presentation.**

| | | | |
|---|---|---|---|
| Cost of Ownership | Tableau is a competitively priced software that is available for a trial download. | Versatility | Multi-purpose package that can be used to build an entire application |
| Big data compatibility | Tableau has become one of the big go-to software programs for Data visualization due to the wide variety of tools it provides and compatibility with Big Data platforms such as Hadoop. | | |



# Know your Tools

*Tableau*

Tableau offers Powerful visualization capabilities, without a single line of code.

Experiment with trend analyses, regressions, correlations.

Scalable, secure and Reliable Cloud and Mobile Connectivity.

## A BUSINESS ANALYSTS' DREAM

Tableau is easy to learn, use, and significantly faster than existing solutions. One can easily see patterns, identify trends and discover visual insights in seconds. No wizards, no scripts.

Tableau facilitates live, up-to-date data analysis that taps into the power of the firm's data warehouse.

Extract data into Tableau's data engine and take advantage of breakthrough in-memory architecture.

## INTEGRATION

Tableau integrates exceptionally well with R and Hadoop, making it a powerful visualization tool for analytics and big data use cases. Developers creating web applications can integrate fully interactive Tableau content into their applications via the JavaScript API.

# Q & A

## Cảm ơn đã theo dõi

Chúng tôi hy vọng cùng nhau đi đến thành công.