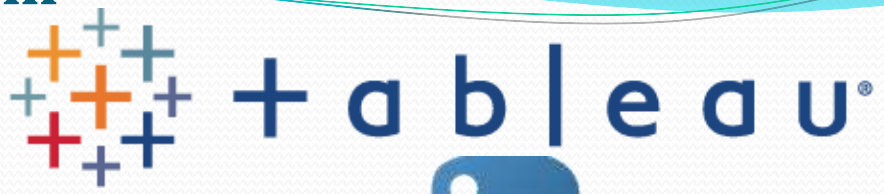
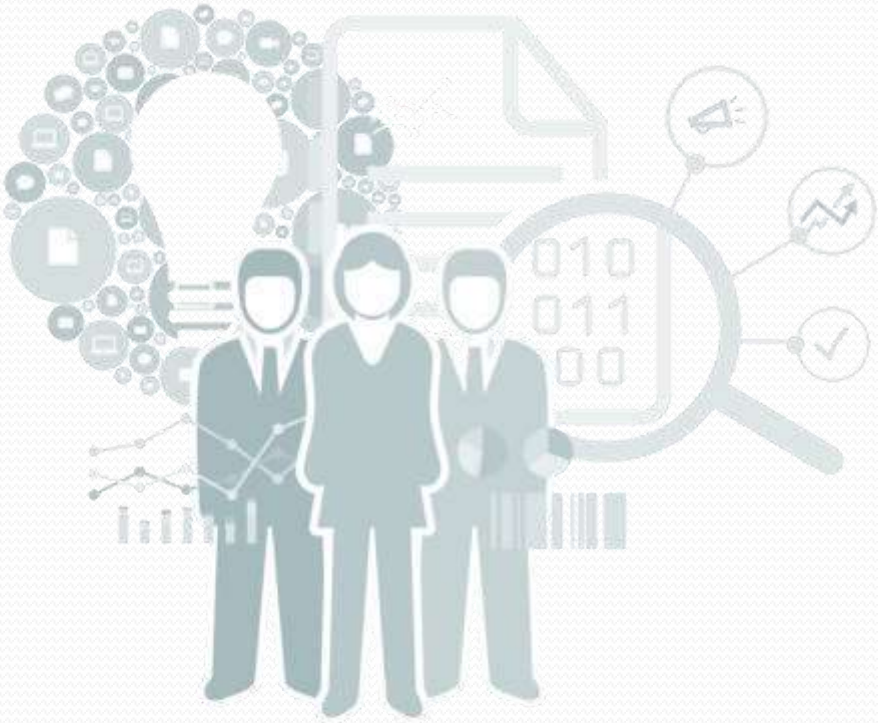
Introduction to HADOOP and HADOOP Architecture

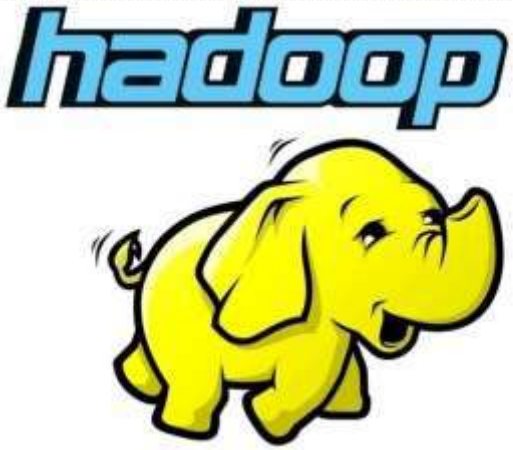
(Unit-III)

Dr. V. K. Patle
Assistant Professor
S. o. S. Computer Science & IT
Pt. Ravishankar Shukla University, Raipur (C.G.)

Tools Covered in Program







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

Why SAS?

#1 Market Leader in Analytics

The largest independent vendor in the business intelligence market

The De facto industry standard for Clinical Data Analysis

INTEGRATED PLATFORM FOR END TO END SOLUTIONS:
SAS provides an integrated set of software products and services and integrated technologies for information management, advanced analytics and reporting.

BUSINESS SOLUTIONS ACROSS DOMAINS AND INDUSTRIES:
Unmatched domain specific industry focused analytics solutions

Used in 60,000+ companies in over 135 countries

"Analytics powerhouse"

The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013

Why R?

Highest Paid IT Skill
Linkedin Skills and O'Reilly Survey, 2016

Most-used data science language after SQL
O'Reilly Survey, Jan 2014

75% of data professionals use R
Rexer Survey, Oct 2015

Second best programming languages for data science
O'Reilly Survey, 2016

Supports close to 10,000 free packages
CRAN Figure as on December 2016

R is the #1 Google Search for Advanced Analytics software
Google Trends, April 2016

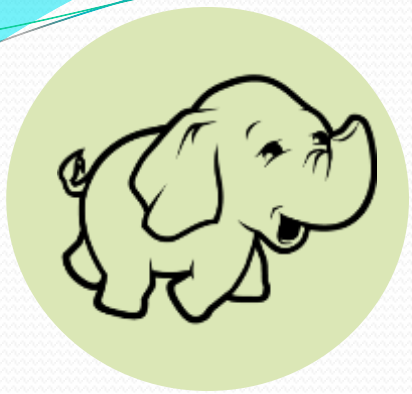
R is #13 of all Programming Languages
Redmonk Language Ratings, June 2015

Demand for R language skills is on the rise.

Companies Already Onboard R

Facebook
Google
Twitter
McKinsey
ANZ Bank
BCG
Uber
Lloyds of London
& Many More…

What is 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.

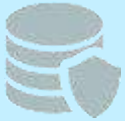
BUSINESS SOLUTIONS ACROSS DOMAINS AND INDUSTRIES:

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

Why Hadoop?

Top 5 Reasons Organizations are using Hadoop

Low Cost	Computing Power	Scalability	Storage Flexibility	Data Protection
				

Enterprises using Hadoop

Top 5 Industries using Hadoop:

- Computer Manufacturing
- Business Services
- Finance
- Retail & Wholesale
- Education & Government





Why Python?

Python is a powerful, flexible, open-source language that is easy to learn, easy to use, and has powerful libraries for data manipulation and analysis

What are the reasons for its sudden popularity?

Cost of Ownership	Python is an open source software that is free to download.
Versatility	Multi-purpose language that can be used to build an entire application
Big data compatibility	Python has become one of the big go-to languages for big data processing due to its wide selection of libraries

Python offers extensive analytics capabilities for Text & Predictive Analytics.

IDLE & Spyder IDE is widely used for data mining.

Big Data Analytics made possible by PyDoop and Scipy

A Data Scientists' Dream

Python is particularly useful in data analytics because it has a rich library for reading and writing data, running calculations on the information and creating graphical representations of data sets.

We can write map reduce programs in python using PyDoop. Here is where Python scores over R. While R uses in-memory processing, Python using PyDoop can process PetaBytes of data

Integration

In industry, the data science trend shows increasing popularity of Python. A Python-based application stack can more easily integrate a data scientist who writes Python code, since that eliminates a key hurdle in productionizing a data scientist's work.

Why Python?

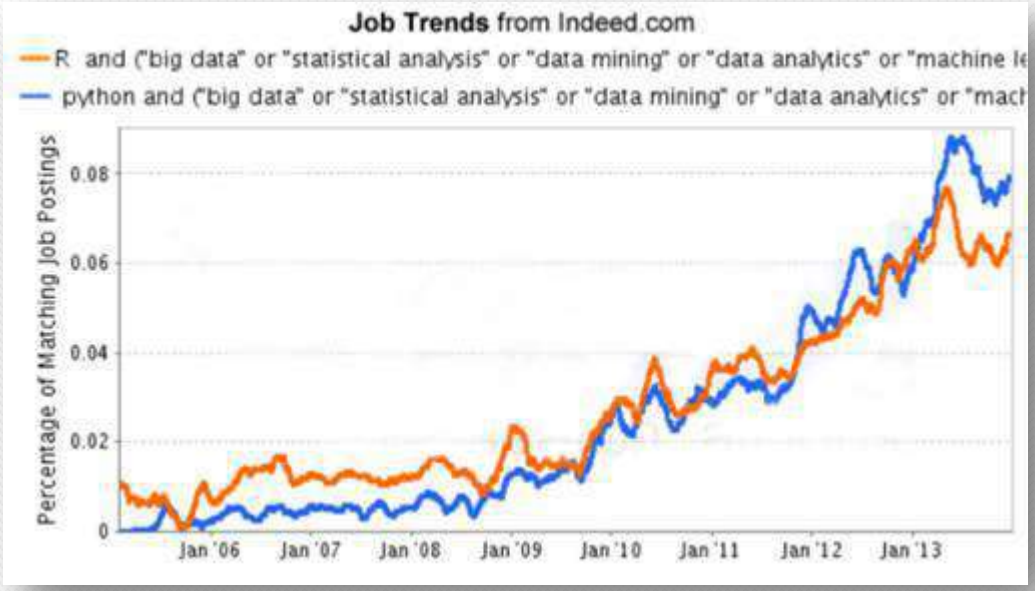
Official language of Google

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

2nd most popular data science language
KDNuggets 2013

Among top in-demand data science skills
KDNuggets, Dec2014

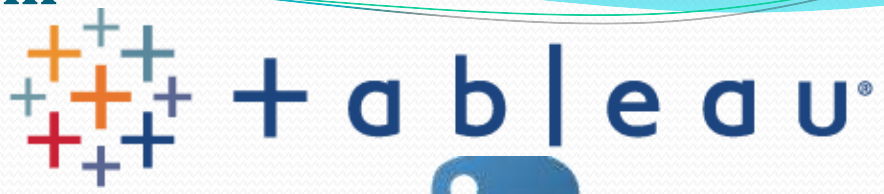
Ranked #1 of all programming languages
Codeeval rankings, Feb 2015

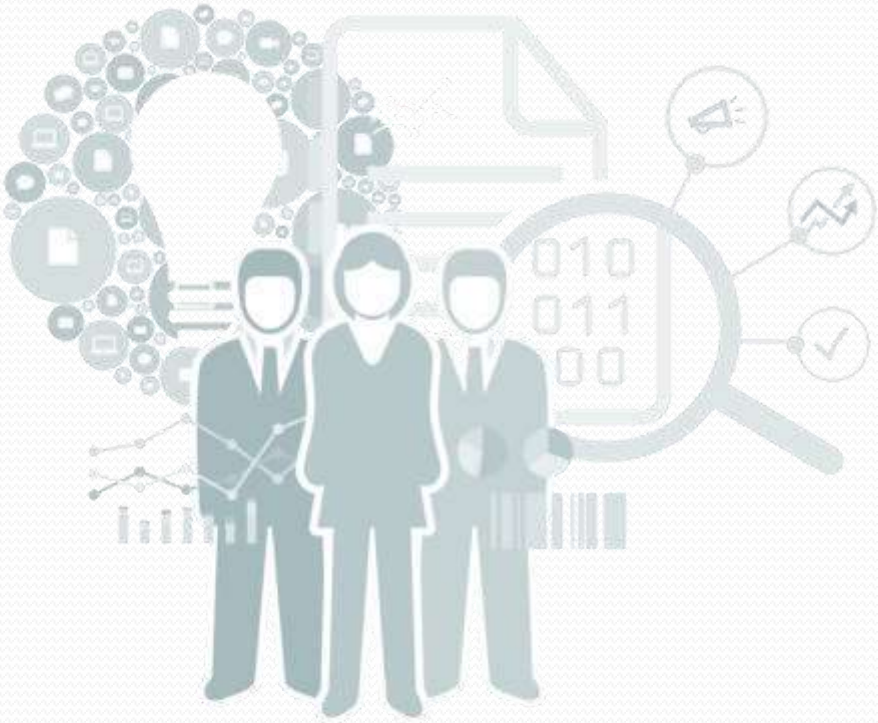


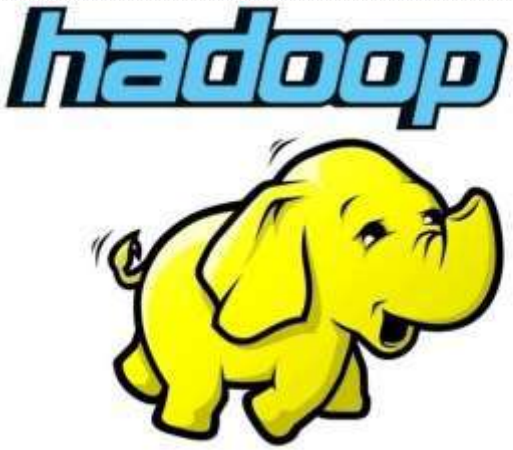
Companies Already Onboard Python

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

Tools Covered in Program







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

Business Drivers and sceneries for large data

- **Social media and websites**
- **IT** — services, Software and Hardware services and support. Finance: Better and deeper understanding of risk to avoid credit crisis
- **Telecommunication**: More reliable network where we can predicate and prevent
- **Media**: More content that is lined up with your personal preferences
- **Life science**: Better targeted medicine with fewer complications and side effects
- **Retail**: A personal experience with product and offer that are just what and you need
- Google, yahoo and others need to index the entire internet and return searched results in milliseconds

Challenges in Big Data Storage and Analysis

- **'Slow to process, can't scale**
  - Disk seek for every access
  - Buffered reads, locality -5 still seeking every disk page
  - It not Storage Capacity but access speeds which is the bottleneck
  - Challenges to both store and analyze datasets
  - Scaling is expensive
- **Hard Drive capacity to process**
  - IDE drive — 75 MB/sec, loms seek
  - DATA drive — 300MB/s, 8.5ms seek
  - SSD — 800MB/s, 2 ms "seek"
  - Apart from this analyze, compute, aggregation, processing dealy etc..
- **Unreliable machines: Risk**
  - Machine 1 time in 3 years mean time between failures
  - 1000 Machines 1 day mean time between failures

Challenges in Big Data Storage and Analysis continues...

- **Reliability**

  Partial failure, graceful decline rather than full halt

  Data recoverability, if a node fails, another picks up its workload

  Node recoverability, a fixed node can rejoin the group without a full group restart

- **Scalability, adding resources adds load capacity**

  Backup

  Not affordable, expensive(faster, more reliability more cost)

  Easy to use and Secure

- **Process data in parallel**

Process data in parallel ? — not simple

An Idea; _Parallelism

Transfer speed improves at a greater rate than seek speed.

Process read/write parallel rather then sequential.

* 1 drive — 75 MB/sec 16 days for 100TB
* 1000 drives — 75 GB/sec 22 minutes for 100TB

A problem: Parallelism is Hard

Synchronization

Deadlock

Limited bandwidth

Timing issues and co-ordination

Spilt & Aggregation

Computer are complicate

Driver failure

Data availability

Co-ordination

Why Not distributed computing

Common Challenges in Distributed computing

We have distributed computing and it also come up with some

- **Resource sharing**. Access any data and utilize CPU resource across the system.
- Portability, reliable,
- **Concurrency**: Allow concurrent access, update of shared resource, availability with high throughput
- **Scalability**: With data, with load
- **Fault tolerance** : By having provisions for redundancy and recovery
- **Heterogeneity:** Different operating system, different hardware
- **Transparency:** Should appear as a whole instead of collection of computers

Hide details and complexity by accomplishing above challenges from the user and need a common unified interface to interact with it.

To address most of these challenges(but not all) Hadoop come in.

Hadoop origins

Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model. It is designed to scale up from single servers to thousands of machines, each providing computation and storage.

Hadoop is an open-source implementation of Google MapReduce, GFS(distributed file system).

Hadoop was created by Doug Cutting, the creator of Apache Lucene, the widely used text search library.

Hadoop fulfill need of common infrastructure

— Efficient, reliable, easy to use

— Open Source, Apache License

The Name 'Hadoop' ?

What's up with the names?
When naming software projects, Doug Cutting seems to have been inspired by his family.
Lucene is his wife's middle name, and her maternal grandmother's first name.
His son, as a toddler, used Nutch as the all-purpose word for meal and later named a yellow stuffed elephant Hadoop.
Doug said he "**was looking for a name that wasn't already a web domain and wasn't trademarked, so I tried various words that were in my life but not used by anybody else. Kids are pretty good at making up words.**"

Hadoop Design Axioms

- Store an process large amounts of data (PetaBytes)
- Performance, storage, processing scale linearly
- Compute should move to data
- Simple core , modular and extensible
- Failure is normal, expected
- Manageable and Heal self
- Design run on commodity hardware-cost effective

Does Hadoop achieves complete parallelism?

For Storage and Distributed computing (MapReduce)

Spilt up the data

- Process Data in parallel
- Sort and combine to get the answer
- Schedule, Process and aggregate independently
- Failures are independent, Handle failures.
- Handle fault tolerance

Hadoop History

+2002-2004 Doug cutting and Mike Cafarella started working on Nutch

• **2003-2004:** Google publishes GFS and MapReduce paper

• **2004 :** Doug cutting adds DFS and Mapreduce support to Nutch

•Yahoo ! Hires Cutting , bulid team to develop Hadoop

#2007: NY time converts 4 TB of archive over 100 EC2 cluster of Hadoop.

•Web scale deployement at Y!,Facebook,twitter.

• **May 2009**: Yahoo does fastest sort of a T B, 62secs over 1460nodes

•Yahoo sort a PB in 16.25hrs over 3658 nodes

Hadoop V/S RDBMS

An Elephant can't jump. But can carry heavy load !!!

• A fundamental tenet of relational databases structure defined by a schema, what about Large data sets are often unstructured or semi-structured, Hadoop is the best choice. Hadoop MR framework uses key/value pairs as its basic data unit, which is flexible enough to work with the less-structured data types.

Scaling commercial relational databases is expensive and limited.

• High-level declarative language like SQL, Block box Query engine. You query data by stating the result you want and let the database engine figure and drive it. you can build complex statistical

Hadoop V/S RDBMS cont..

models from your data or analytical reporting or reformat your image data. SQL is not well designed for such tasks. MapReduce tries to collocate the data with the compute node, so data access is fast since it is local.

- Coordinating the processes in a large-scale distributed computation is a challenge. HDFS and MR made easy split, store. process and aggregate.
- To run a bigger database you need to buy a bigger machine. the high-end machines are not cost effective for many applications. For example, a machine with four times the power of a standard PC costs a lot more than putting four such PCs in a cluster. Hadoop is designed to be a scale-out architecture operating on a cluster of commodity hardware . Adding more resources means adding more machines to the Hadoop cluster.

Hadoop V/S RDBMS cont..

Effective cost per user TB: $250/TB

Other solutions(RDBMS) cost in the range of $100 to $100K per user TB

Hardest aspect is gracefully handling partial failure— when you don't know if a remote process has failed or not—and still making progress with the overall computation. MapReduce spares the programmer from having to think about failure, since the implementation detects failed map or reduce tasks and reschedules with suitable replacements . MapReduce is able to do this since it is a shared-nothing architecture, meaning that tasks have no dependence on one other.

Hadoop V/S RDBMS cont..

Hardware failure: as soon as you start using many pieces of hardware, the chance that one will fail is fairly high. A common way of avoiding data loss is through replication. Redundant copies of the data are kept by the system so that in the event of failure, there is another copy available. Node failure and disk
failure efficient handle in Hadoop frame work.

	Traditional RDBMS	MapReduce
Data size	Gigabytes	Petabytes
Access	Interactive and batch	Batch
Updates	Read and write many times	Write once, read many times
Structure	Static schema	Dynamic schema
Integrity	High	Low
Scaling	Nonlinear	Linear

Hadoop V/S RDBMS cont..

Is Hadoop alternative for RDBMS ?

Hadoop is not replacing the traditional data systems used for building analytic applications — the RDBMS, EDW and MPP systems — but rather is a complement.

- Interoperate with existing systems and tools, at the moment Apache Hadoop is not a substitute for a database
- No Relation, Key Value pairs
- Big Data, unstructured (Text) & semi structured (Seq / Binary Files) Structured (H base-Google Big Table)Works fine together with RDBMs, Hadoop is being used to large quantities of data into something more manageable.

Why SAS?

#1 Market Leader in Analytics

The largest independent vendor in the business intelligence market

The De facto industry standard for Clinical Data Analysis

INTEGRATED PLATFORM FOR END TO END SOLUTIONS:
SAS provides an integrated set of software products and services and integrated technologies for information management, advanced analytics and reporting.

BUSINESS SOLUTIONS ACROSS DOMAINS AND INDUSTRIES:
Unmatched domain specific industry focused analytics solutions

Used in 60,000+ companies in over 135 countries

"Analytics powerhouse"

The Forrester Wave™: Big Data Predictive Analytics Solutions, Q1 2013

Why R?

Highest Paid IT Skill
Linkedin Skills and O'Reilly Survey, 2016

Most-used data science language after SQL
O'Reilly Survey, Jan 2014

75% of data professionals use R
Rexer Survey, Oct 2015

Second best programming languages for data science
O'Reilly Survey, 2016

Supports close to 10,000 free packages
CRAN Figure as on December 2016

R is the #1 Google Search for Advanced Analytics software
Google Trends, April 2016

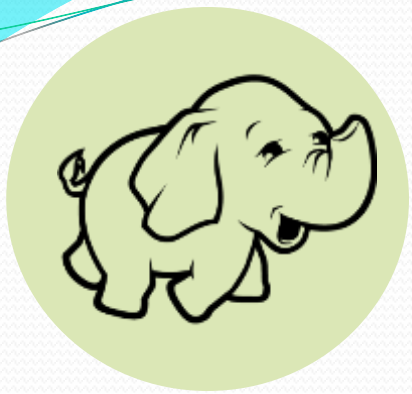
R is #13 of all Programming Languages
Redmonk Language Ratings, June 2015

Demand for R language skills is on the rise.

Companies Already Onboard R

Facebook
Google
Twitter
McKinsey
ANZ Bank
BCG
Uber
Lloyds of London
& Many More…

What is 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

Feature of Hadoop?

Top 5 Reasons Organizations are using Hadoop

Low Cost	Computing Power	Scalability	Storage Flexibility	Data Protection

Enterprises using Hadoop

Top 5 Industries using Hadoop:

- Computer Manufacturing
- Business Services
- Finance
- Retail & Wholesale
- Education & Government

Why Python?

Python is a powerful, flexible, open-source language that is easy to learn, easy to use, and has powerful libraries for data manipulation and analysis

What are the reasons for its sudden popularity?

Cost of Ownership	Python is an open source software that is free to download.
Versatility	Multi-purpose language that can be used to build an entire application
Big data compatibility	Python has become one of the big go-to languages for big data processing due to its wide selection of libraries

Python offers extensive analytics capabilities for Text & Predictive Analytics.

IDLE & Spyder IDE is widely used for data mining.

Big Data Analytics made possible by PyDoop and Scipy

A Data Scientists' Dream

Python is particularly useful in data analytics because it has a rich library for reading and writing data, running calculations on the information and creating graphical representations of data sets.

We can write map reduce programs in python using PyDoop. Here is where Python scores over R. While R uses in-memory processing, Python using PyDoop can process PetaBytes of data

Integration

In industry, the data science trend shows increasing popularity of Python. A Python-based application stack can more easily integrate a data scientist who writes Python code, since that eliminates a key hurdle in productionizing a data scientist's work.

Why Python?

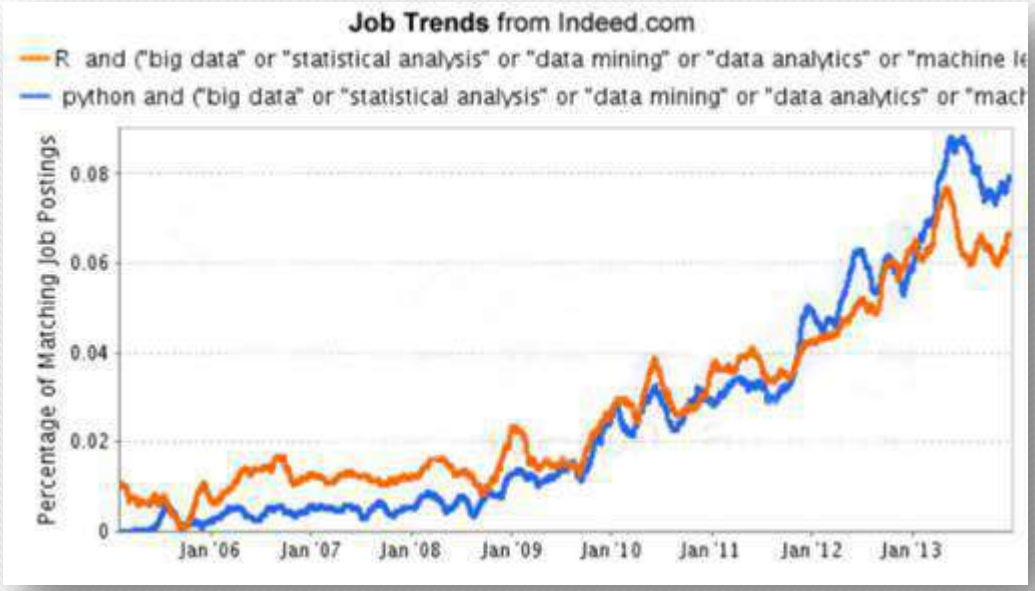
Official language of Google

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

2nd most popular data science language
KDNuggets 2013

Among top in-demand data science skills
KDNuggets, Dec2014

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



Companies Already Onboard Python

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

References:-

1.https://www.techopedia.com/

2. https://pediaa.com/

3.https://www.geeksforgeeks.org/

THANKYOU