Distributed Data Mining Lab

Meet the team

DataEngineers


Chaitanya Aggarwal


Ishmeet Kaur


Munwar Ahmad


Roba Abbas

Goals

Learn Distributed Technologies

ANALYZE Frameworks e.g Spark,Hadoop etc.

Work real application on distributed platform

Common Scenario in companies

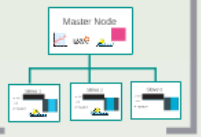
Why Hadoop?

Why Spark?

Technologies Used

Project Architecture 1

Project Architecture 2



Project Architecture 1

Weekly Milestones

Obstacles

Biggest Error

Conclusion

- Introduction to distributed technologies.
- Setting up the cluster from the scratch.
- Distributed programming experience.
- Implemented and Visualized the protein locations.
- Mini talks.

Learning curve

Visualizing data with

https://goo.gl/7AdbEz

Parsing and storing in Elastic Search

Updating Protein Localization

Goals

- LEARN Distributed Technologies
- ANALYZE Frameworks e.g Spark,Hadoop etc.
- IMPLmet application on distributed platform

Common Scenario in companies

Why Hadoop?

Meet the team

DataEngineers


Chaitanya Aggarwal


Ishmeet Kaur

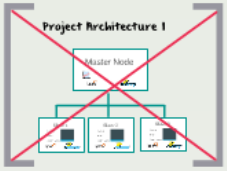

Muneer Ahmad


Roba Abbas

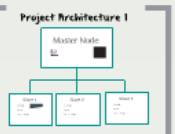
Distributed Data Mining Lab

Why Spark?

Project Architecture 1

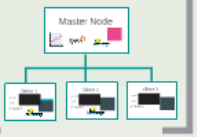


Project Architecture 1



Technologies Used

Project Architecture 2

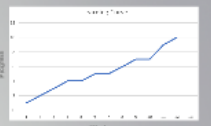


Weekly Milestones

Obstacles

Biggest Error

Learning curve



Visualizing data with

https://goo.gl/7AdbEz

Parsing and storing in Elastic Search

Updating Protein Localization

Conclusion

- Introduction to distributed technologies.
- Setting up the cluster from the scratch.
- Distributed programming experience.
- Implemented and Visualized the protein locations.
- Mini talks.

Distributed Data Mining Lab

Meet the team

DataEngineers

Chaitanya Aggarwal

Ishmeet Kaur

Muneer Ahmad

Roba Abbas

Goals

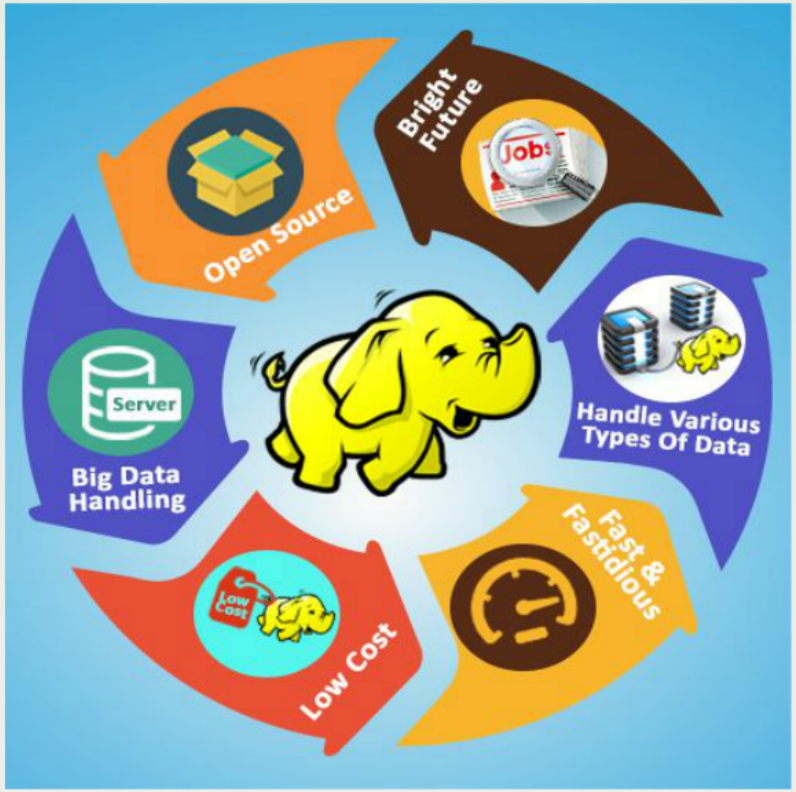
LEARN Distributed Technologies

ANALYZE Frameworks e.g Spark,Hadoop etc.

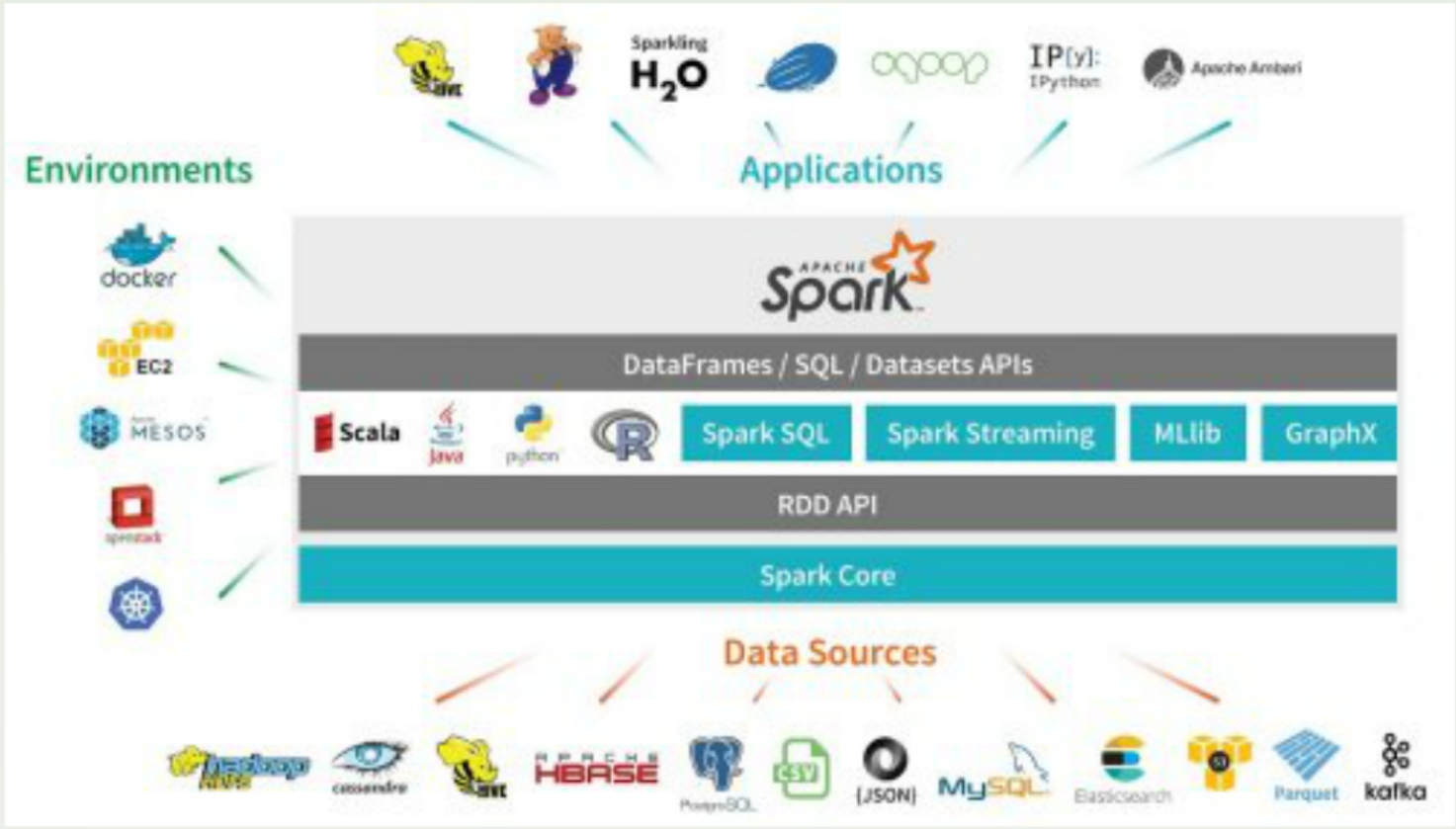
RUN real application on distributed platform

Common Scenario in companies

Why Hadoop?

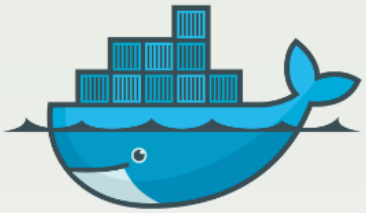

Open Source
Bright Future
Job!
Handle Various Types Of Data
Fast & Fastidious
Low Cost
Low Cost
Big Data Handling
Server

Why Spark?


Environments
Applications
docker
EC2
MESOS
Sparkling H2O
IP[y]: IPython
Apache Ambari
APACHE Spark
DataFrames / SQL / Datasets APIs
Scala
Java
python
Spark SQL
Spark Streaming
MLlib
GraphX
RDD API
Spark Core
Data Sources
HBASE
PostgreSQL
CSV
{JSON}
MySQL
Elasticsearch
Parquet
kafka

Technologies Used

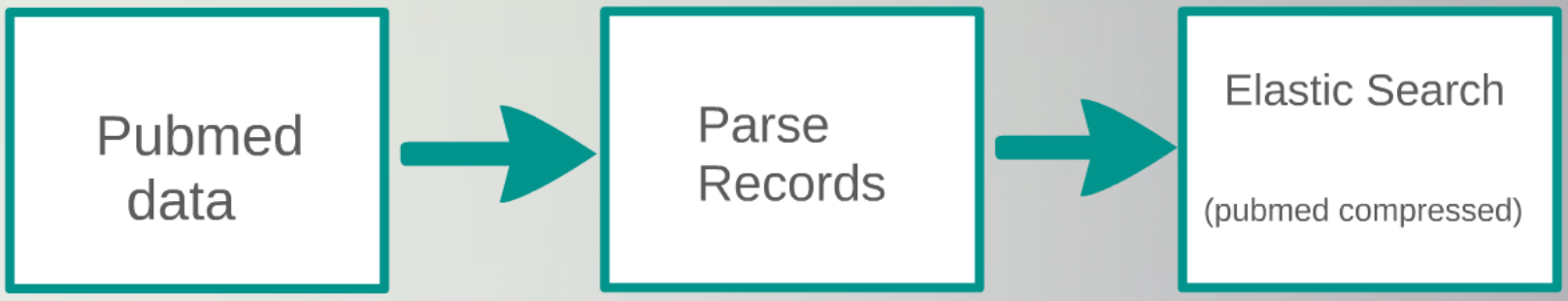
elasticsearch



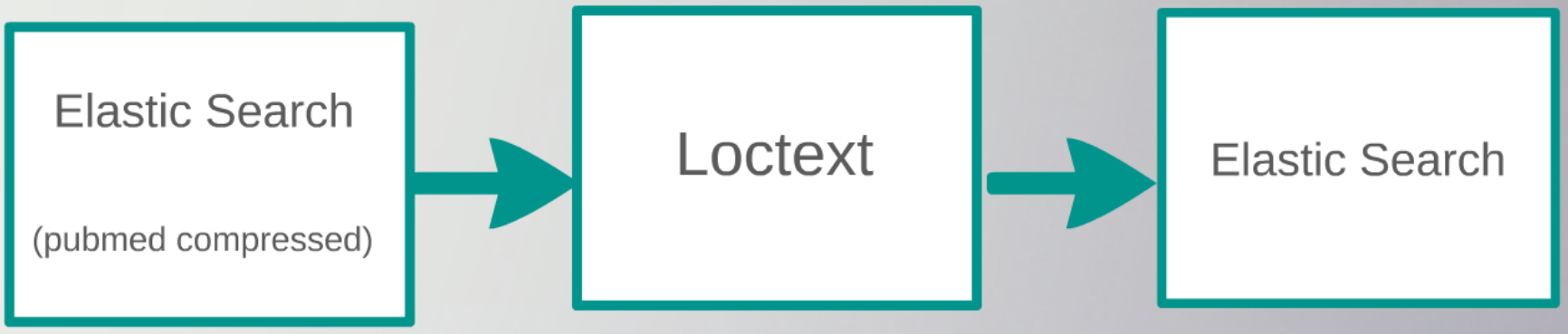
MLlib

kibana

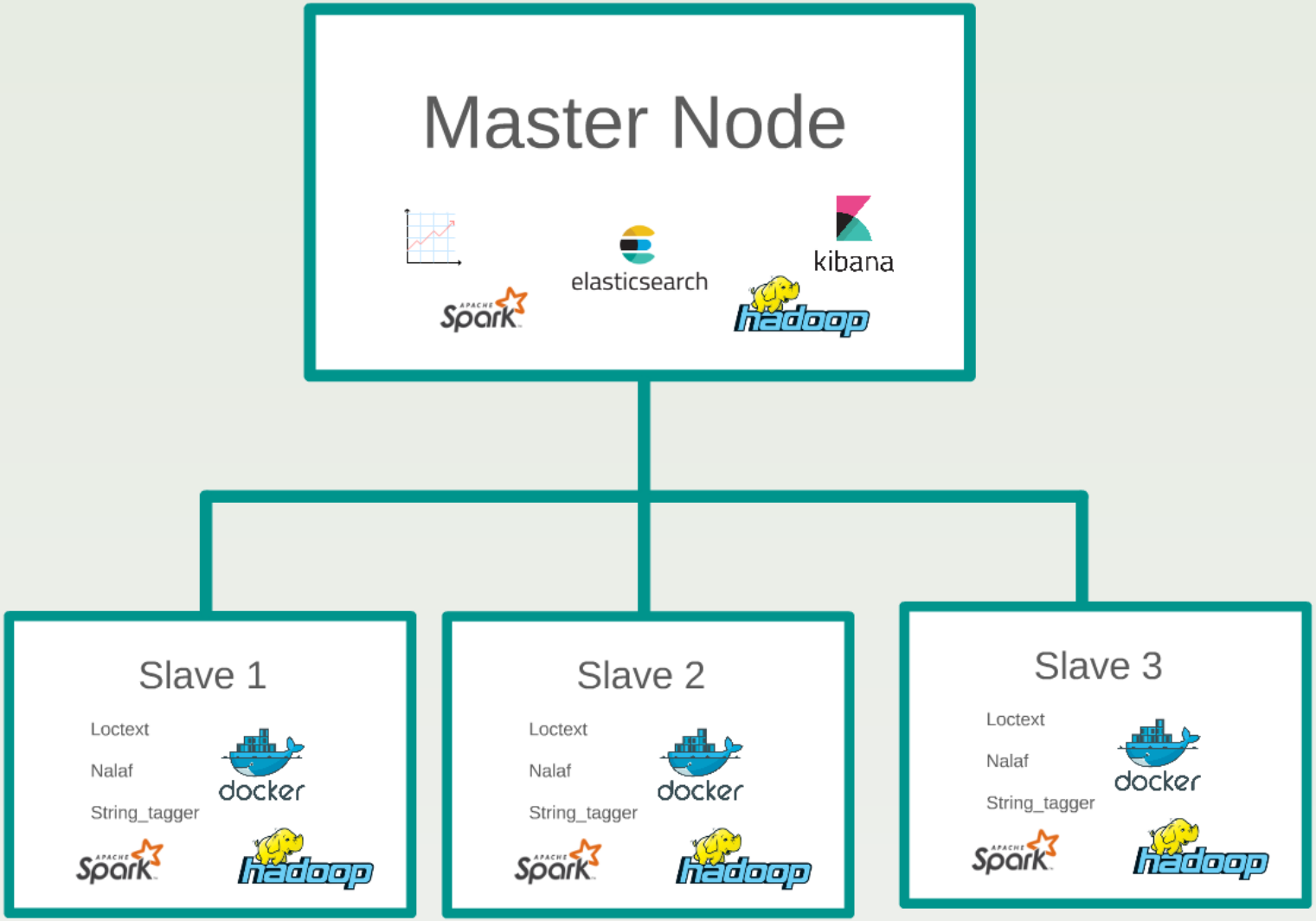
Parsing and storing in Elastic Search



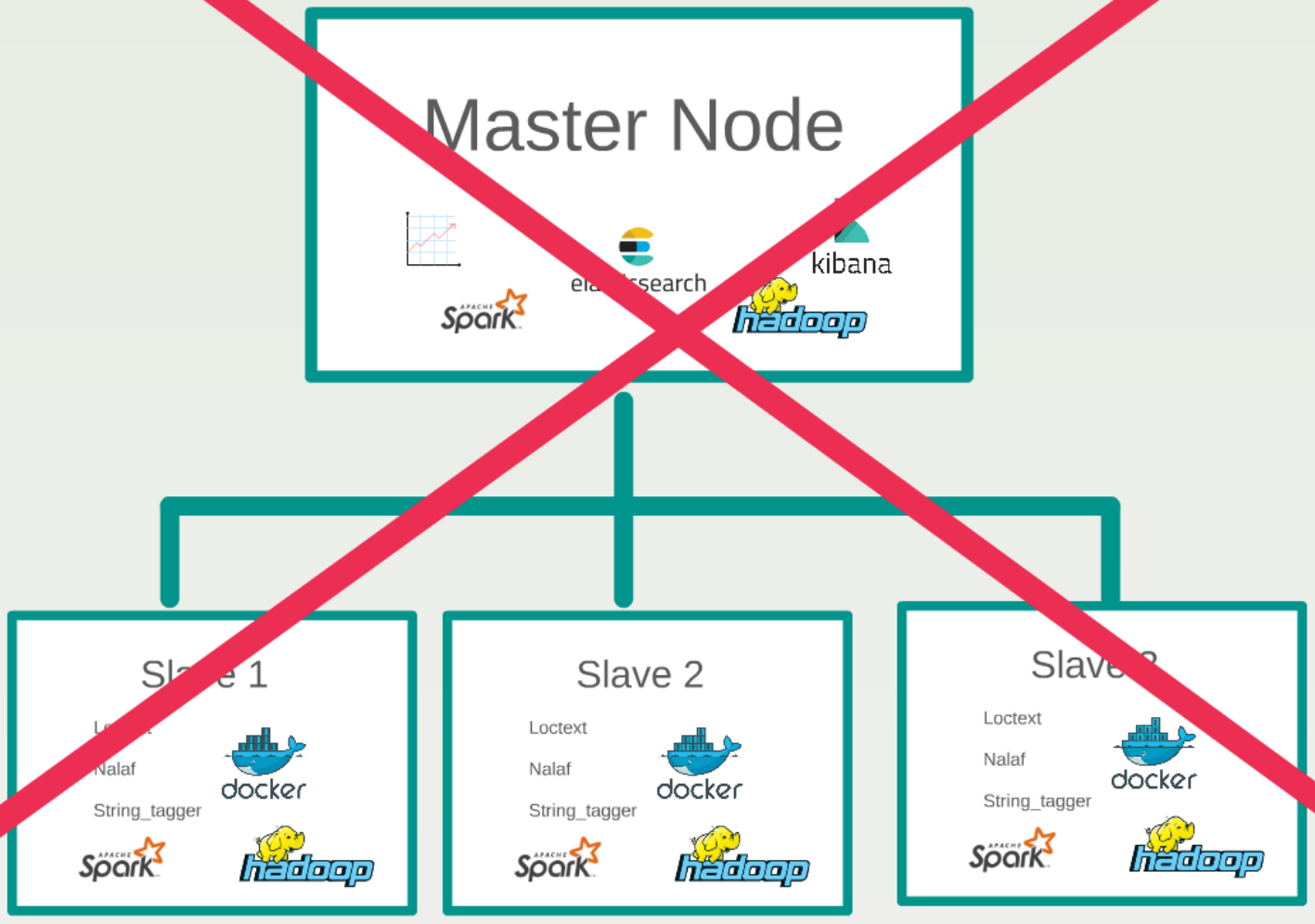
Updating Protein Localization



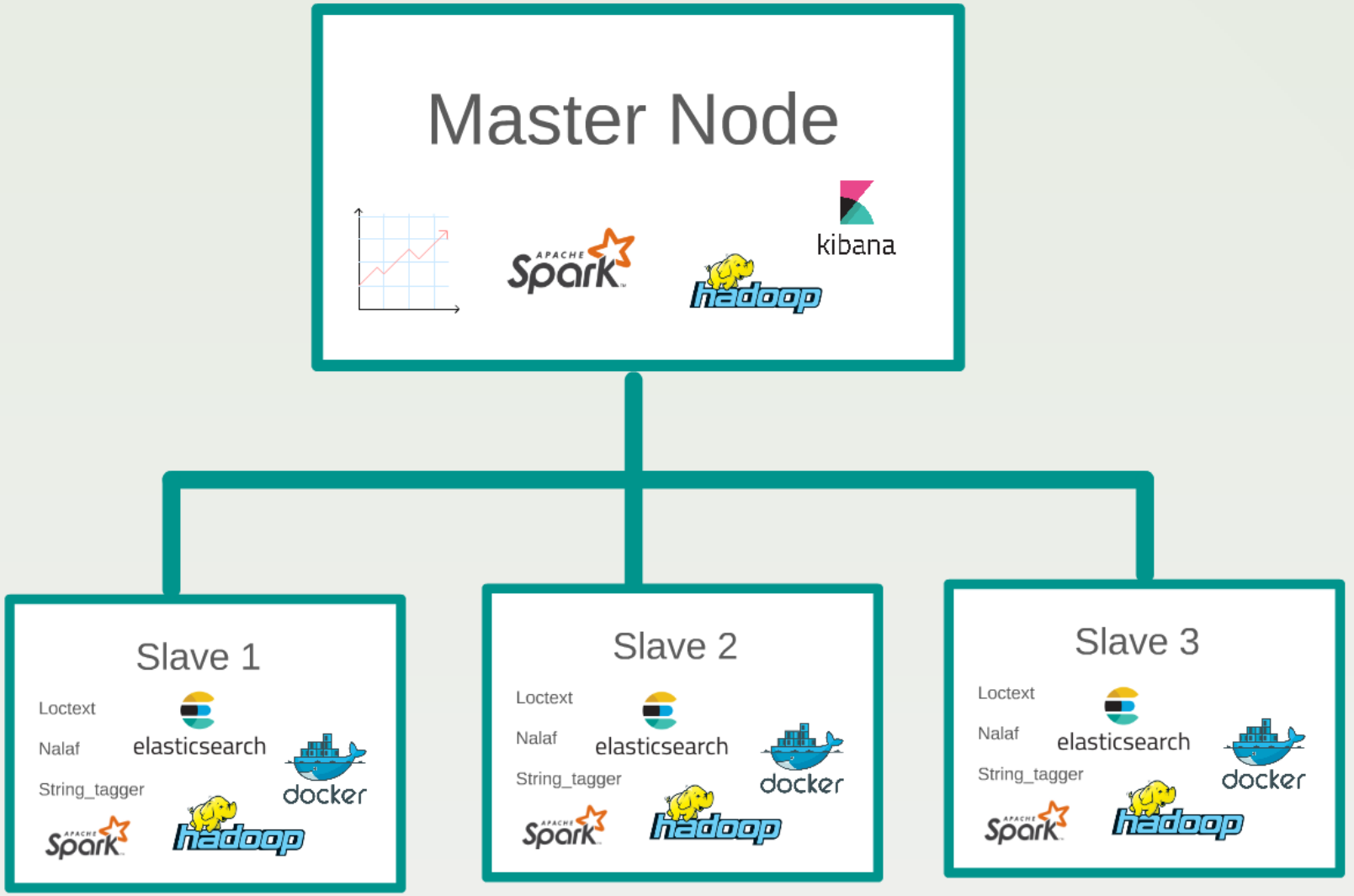
Project Architecture 1



Project Architecture 1



Project Architecture 2



Visualizing data with

https://goo.gl/7AdbEz

Weekly Milestones

Virtual Machines setup

Single and MultiNode Cluster setup

Exploration and programming of Hadoop & Spark

Extraction of NCBI Database and Elsevier Data

Elastic Search, LocText, Parsing and storage of Full Text Papers

Text-mine the relationship of Protein & Cell with LocText and MapReduce

Obstacles

Port blocking / Live nodes
Java heap space

elasticsearch

Unstructured data structure.
Memory error

HTTPs error Efetch API

LocText

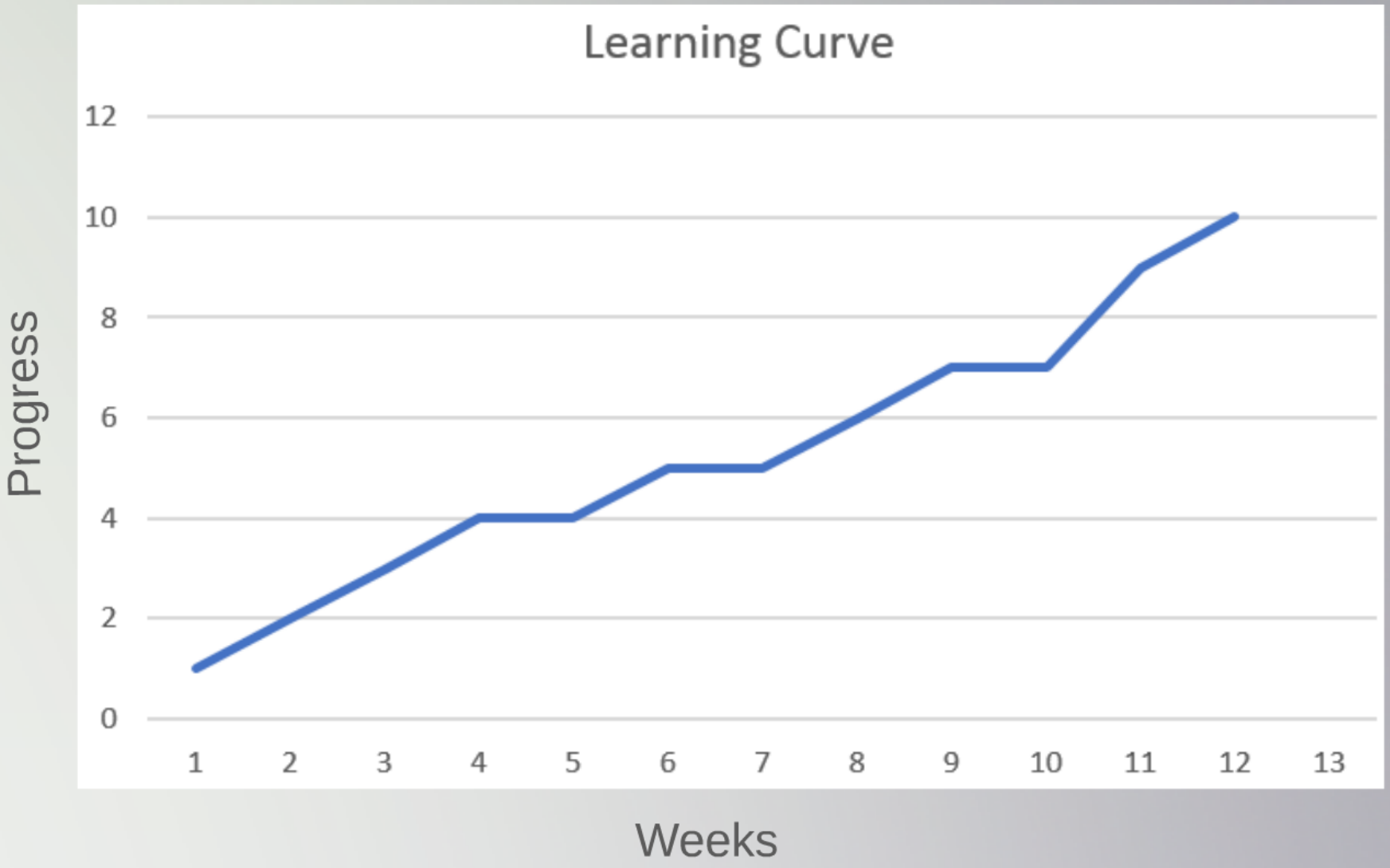
Import Errors (String Tagger)
Input to Loctext is not a protein

Docker IO
Docker ip tables error

Biggest Error

```
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
@    WARNING: REMOTE HOST IDENTIFICATION HAS CHANGED!     @
@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@@
IT IS POSSIBLE THAT SOMEONE IS DOING SOMETHING NASTY!
Someone could be eavesdropping on you right now (man-in-the-middle attack)!
It is also possible that a host key has just been changed.
The fingerprint for the ECDSA key sent by the remote host is
SHA256:MoxgOTJN4TImrj2Mdqyz16XWnpr0WZlSA+gBc2LFGEk.
Please contact your system administrator.
```

Learning curve



Conclusion

- Introduction to distributed technologies.
- Setting up the cluster from the scratch.
- Distributed programming experience.
- Implemented and Visualized the protein locations.
- Mini talks.

Goals

- LEARN Distributed Technologies
- ANALYZE Frameworks e.g Spark,Hadoop etc.
- IMPLmet application on distributed platform

Common Scenario in companies

Why Hadoop?

Why Spark?

Meet the team

DataEngineers


Chaitanya Aggarwal

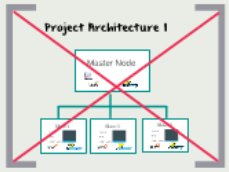

Ishmeet Kaur


Muneer Ahmad

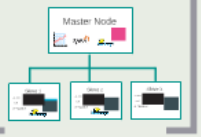

Roba Abbas

Distributed Data Mining Lab

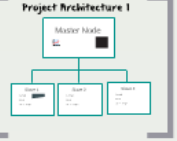
Project Architecture 1



Project Architecture 2



Project Architecture 1



Technologies Used

Weekly Milestones

Obstacles

Biggest Error

Visualizing data with

https://goo.gl/7AdbEz

Learning curve

Parsing and storing in Elastic Search

Updating Protein Localization

Conclusion

- Introduction to distributed technologies.
- Setting up the cluster from the scratch.
- Distributed programming experience.
- Implemented and Visualized the protein locations.
- Mini talks.