Addis Ababa University
አዲስ፡አበባ፡ዩኒቨርሲቲ
SEEK WISDOM, ELEVATE YOUR INTELLECT AND SERVE HUMANITY !

Institute of Geophysics, Space Science, and Astronomy
Unit of Geodesy, Geomatics and Gravimetry
PhD in Remote Sensing and Geospatial Information Science

Course Name: Geospatial Big Data Analytics

Apache Hive: A Warehousing Solution Over a Map-Reduce Framework

Presenter: Tesfahun Endalew

Group Member: Tesfayesus Yimenu and Samson Warkaye (PhD Student)

November, 2023

Presentation One Outline

- Introduction
- History of Apache Hive
- Architecture of Hive
- Data Flow in Hive
- Hive Data Modeling
- Hive Data Types
- Features of Hive
- Advantage and Disadvantages of Hive
-

Introduction

Apache Hive

What is Apache Hive?

- The Apache Hive ™ data warehouse software facilitates reading, writing, and managing large datasets residing in distributed storage using SQL (Standard Query Language).
- Apache hive is a data warehousing tool built on top of Hadoop and used for extracting meaningful information from data. Data warehousing is all about storing all kinds of data generated from different sources at the same location.

Cont…

- The data is mostly available in 3 forms i.e. structured (SQL database), semi-structured(XML or JSON) and unstructured(music or video).
- To process the structured data available in the tabular format we use Hive on top of Hadoop. The Hive is so powerful that it can query Petabytes (PB) of data very efficiently.
- If you have had a look at the Hadoop Ecosystem, you may have noticed the yellow elephant trunk logo that says HIVE, but do you know what Hive is all about and what it does? At a high level, some of Hive's main features include querying and analyzing large datasets stored in HDFS.

Cont…

- It supports easy data summarization, ad-hoc queries, and analysis of vast volumes of data stored in various databases and file systems that integrate with Hadoop. In other words, in the world of big data, Hive is huge.

- **let's start by understanding why Hive came into existence.**

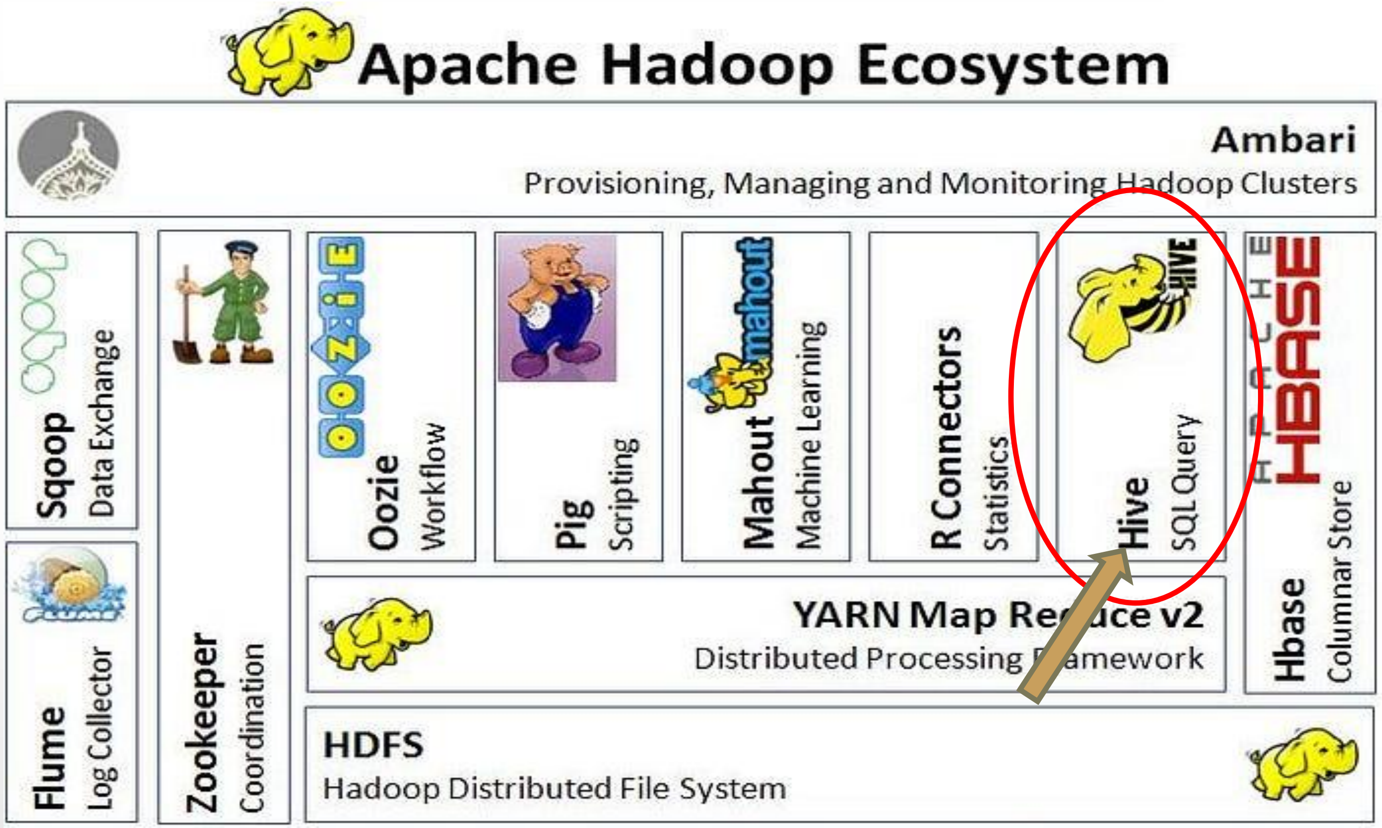


Fig1. ***Apache Hive locations in Apache Hadoop ecosystem***

History of Apache Hive

- Hive has a fascinating history related to the world's largest social networking site that is Facebook.
- Facebook adopted the Hadoop framework to manage their big data. As we know that big data is nothing but massive amounts of data that cannot be stored, processed, and analyzed by traditional systems.
- Hadoop uses Map-Reduce to process data. With Map-Reduce, users were required to write **long** and extensive Java code.
- Not all users were well-versed with Java and other coding languages. Users were comfortable with writing queries in SQL (Structured Query Language), and they wanted a language similar to SQL.
- Enter the HiveQL language. The idea was to incorporate the concepts of tables and columns, just like SQL.

Cont…

- Hive is a data warehouse system that is used to query and analyze large datasets stored in the HDFS. Hive uses a query language called HiveQL, which is similar to SQL.
- As seen from the image below, the user first sends out the Hive queries. These queries are converted into Map-Reduce tasks, and that accesses the Hadoop Map-Reduce system.

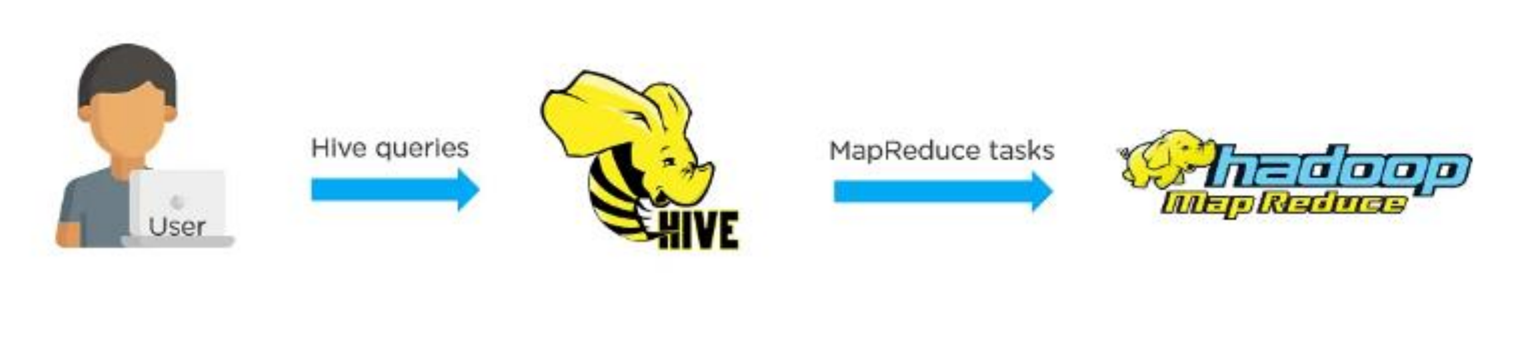


Fig 2 Apache Hive Process

- **let's now take a look at the architecture of the Hive.**

Architecture of Hive

- We start with the Hive client, who could be the programmer who is proficient in SQL, to look up the data that is needed.
- When we look at the structure of Apache Hive On Hadoop, we must not forget to include the *UI, Driver, Compiler, Metastore,* *and* *Execution Engine* as key components.
- The Hive client supports different types of client applications in different languages to perform queries. Thrift is a software framework. The Hive Server is based on Thrift, so it can serve the requests from all of the programming languages that support Thrift.

The architecture of the Hive is as shown below.

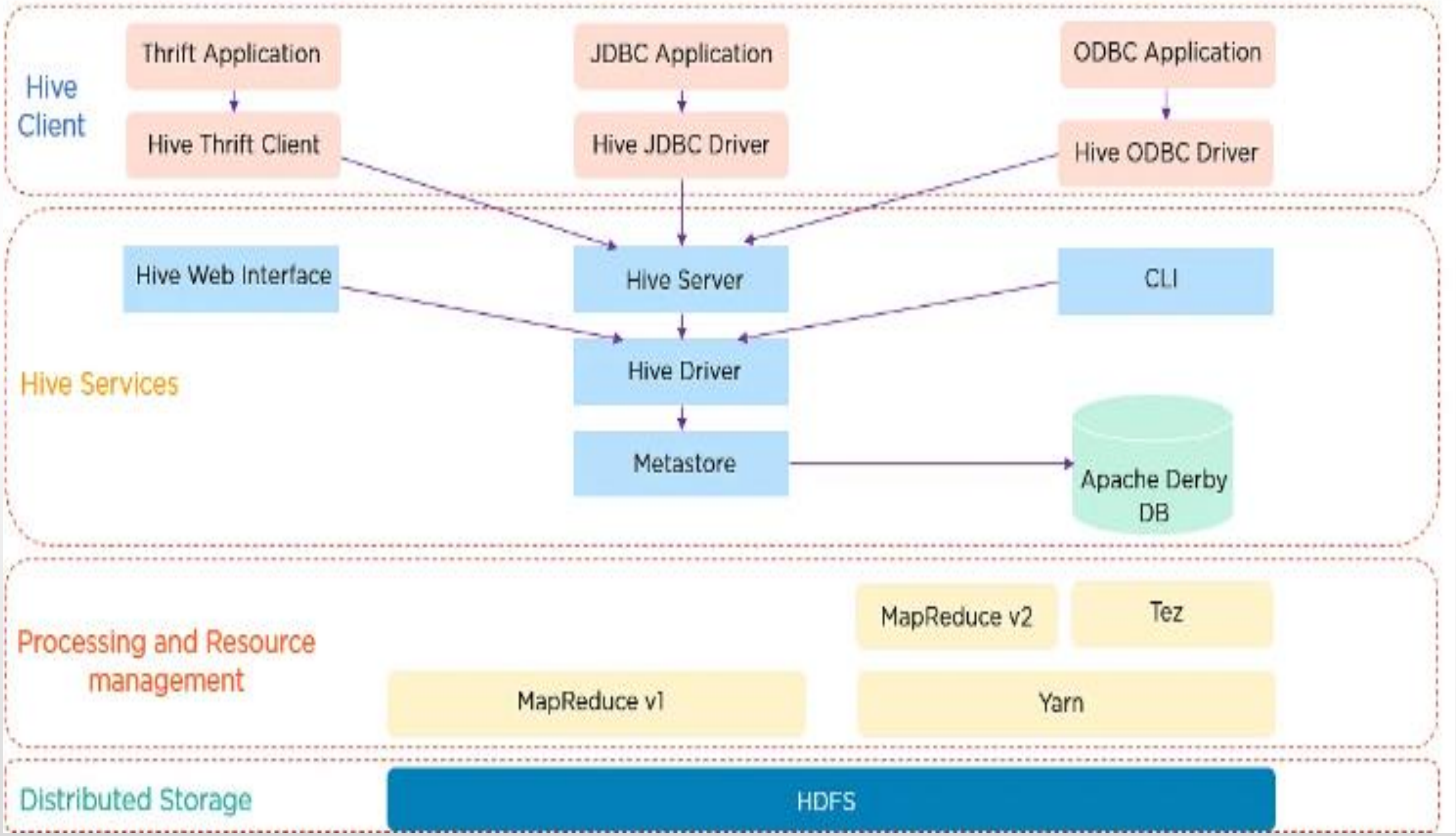


Fig 3 Architecture of Hive

Cont….

- Next, we have the JDBC (Java Database Connectivity) application and Hive JDBC Driver.
- The JDBC application is connected through the JDBC Driver.
- Then we have an ODBC (Open Database Connectivity) application connected through the ODBC Driver.
- All these client requests are submitted to the Hive server.
- In addition to the above, we also have the Hive web interface, or GUI, where programmers execute Hive queries. Commands are executed directly in CLI. Up next is the Hive driver, which is responsible for all the queries submitted.

Cont…

- It performs three steps internally:

1. **Compiler** - The Hive driver passes the query to the compiler, where it is checked and analyzed
2. **Optimizer** - Optimized logical plan in the form of a graph of Map-Reduce and HDFS tasks is obtained
3. **Executor** - In the final step, the tasks are executed

- **Metastore** is a repository for Hive metadata. It stores metadata for Hive tables, and you can think of this as your schema.
- This is located on the Apache Derby DB. Hive uses the Map-Reduce framework to process queries.
- Finally, we have distributed storage, which is HDFS.

Data Flow in Hive

- Data flow in the Hive contains the Hive and Hadoop system.
- Underneath the user interface, we have driver, compiler, execution engine, and metastore.
- All of that goes into the Map-Reduce and the Hadoop file system.

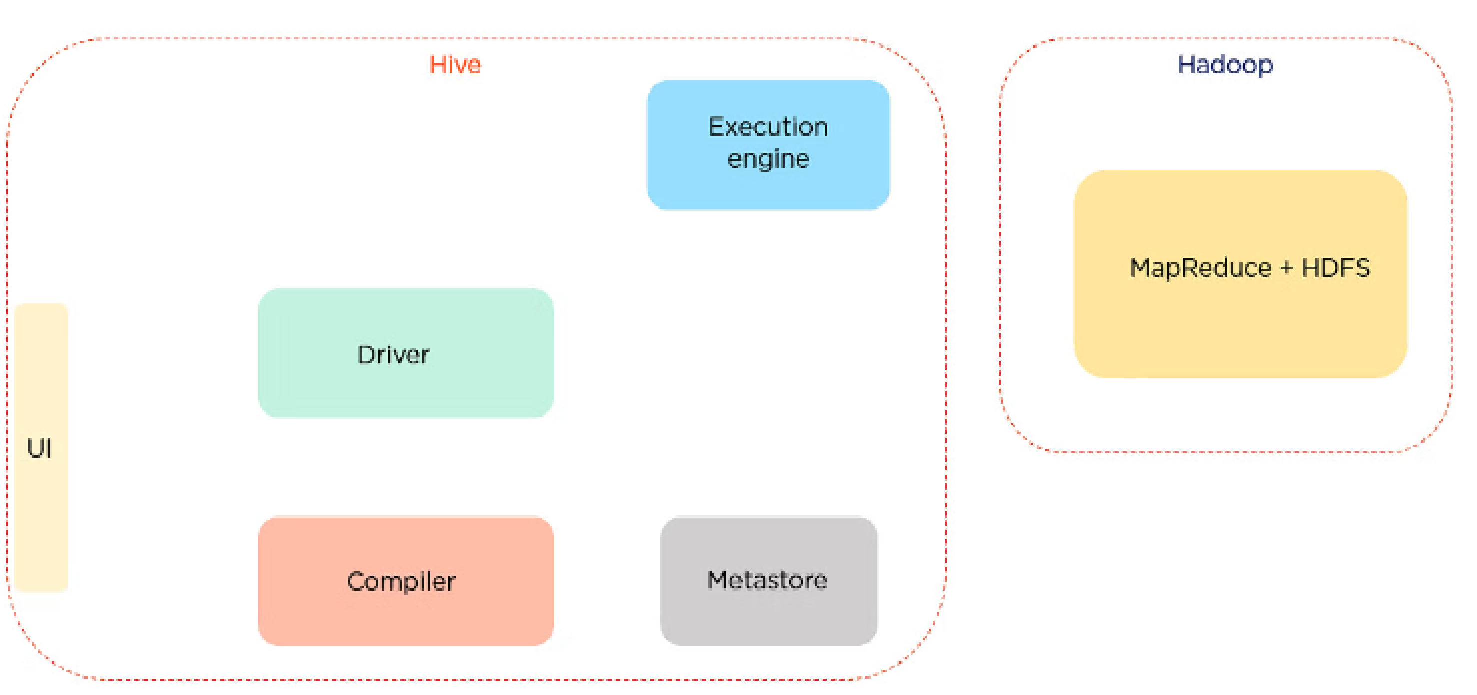


Fig 4 Data flow in Hive

The data flow in the following sequence:

1. We execute a query, which goes into the driver
2. Then the driver asks for the plan, which refers to the query execution
3. After this, the compiler gets the metadata from the metastore
4. The metastore responds with the metadata
5. The compiler gathers this information and sends the plan back to the driver
6. Now, the driver sends the execution plan to the execution engine
7. The execution engine acts as a bridge between the Hive and Hadoop to process the query
8. In addition to this, the execution engine also communicates bidirectional with the metastore to perform various operations, such as create and drop tables
9. Finally, we have a bidirectional communication to fetch and send results back to the client

Cont…

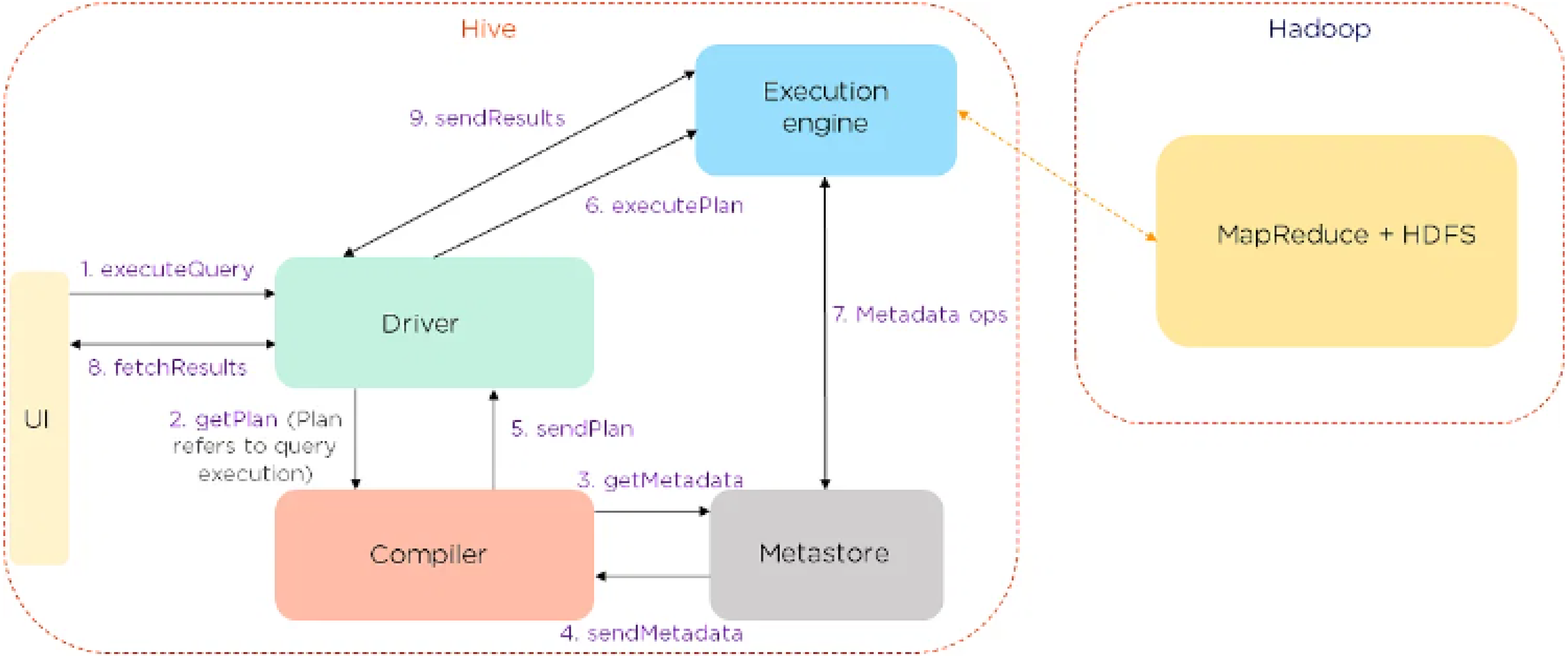


Fig. 5 Data flow in Hive with Sequence

Apache Hive: a Warehousing Solution Over a Map-Reduce Framework

Hive Data Modeling

Hive data modeling, which consists of tables, partitions, and buckets:

1. **Tables:** Tables in Hive are created the same way it is done in RDBMS
2. **Partitions:** Here, tables are organized into partitions for grouping similar types of data based on the partition key
3. **Buckets:** Data present in partitions can be further divided into buckets for efficient querying

Hive Data Types

These are classified as primitive and complex data types.

1. Primitive Data Types:
   - *Numeric Data types- Data types like integral, float, decimal*
   - *String Data type- Data types like char, string*
   - *Date/ Time Data type- Data types like timestamp, date, interval*
   - *Miscellaneous Data type- Data types like Boolean and binary*
2. Complex Data Types:
   - *Arrays- A collection of the same entities. The syntax is: array<data_type>*
   - *Maps- A collection of key-value pairs and the syntax is map<primitive_type, data_type>*
   - *Structs- A collection of complex data with comments. Syntax: struct<col_name : data_type [COMMENT col_comment],…..>*
   - *Units- A collection of heterogeneous data types. Syntax: uniontype<data_type, data_type,..>*

Features of Hive

Now that we have tried to present about the architecture of the Hive, the different data types of Hive, and Hive data modeling, let us look into the Some Apache Hive's features:

1. **Storage:** *Hive supports users to access files from* ***HDFS,*** *Apache HBase, Amazon S3, etc.*
2. **Capable:** *Hive is capable to process very large datasets of Petabytes in size.*
3. **Supported Computing Engine:** *Hive supports Map-Reduce, Tez, and Spark computing engine.*
4. **Framework:** *Hive is a stable batch-processing framework built on top of the Hadoop Distributed File system and can work as a data warehouse.*
5. **Easy To Code:** *Hive uses HIVE query language to query structure data which is easy to code. The 100 lines of java code we use to query a structure data can be minimized to 4 lines with HQL.*
6. **Declarative:** *HQL is a declarative language like SQL means it is non-procedural.*
7. **Drivers:** *JDBC/ODBC drivers are also available in Hive………e.t.c*

Some Advantages of Apache Hive

Advantages	Clarification
Speed	Apache Hive is designed to quickly handle petabytes of data using batch processing. In batch processing, the data is divided as bits and analyzed separately.
Cost	Apache Hive is a cost effective option if your organization's prime focus is on profit. It provides much cheaper options for big data analysis.
Reliability	For Big-data analysis the reliability provided by Apache Hive is far superior than other software solutions. The big-data is replicated each time when it is analyzed. Even in the case of machine malfunctioning, there is no data loss.
Efficiency	Hive is easy to use application for both beginners and experts in programming. Anyone who is familiar with SQL can work with Hive easily. And for projects involving complex coding, Hive allows to divide the work.
Customer Support	Hive includes an attractive customer support service. Their team consists of members who are ready to respond customer queries. Besides that, they ensure Hive is modified with necessary improvements as to improve customer experience.

Some Limitation of Apache Hive

Limitation	Clarification
Does not support OLTP	*Apache Hive doesn't support online transaction processing (OLTP) but Online Analytical Processing(OLAP) is supported.*
Doesn't support subqueries	*Subqueries are not supported.*
Latency	*The latency in the apache hive query is very high.*
Only non-real or cold data is supported	*Hive is not used for real-time data querying since it takes a while to produce a result.*
Transaction processing is not supported	*HQL does not support the Transaction processing feature.*

Thank you for your Attentions!

Next Presenter

Tesfayesus Yimemu

On Apache **Splunk**

splunk>

splunk>

FLUNKING meaning: **to explore natural caves**

splunk:- the “information caves” that organizations struggled with

By caves, its creators meant Big Data and machine-generated data.

American company started in 2003

“Splunk” is a big data platform that simplifies the task of collecting and managing massive volumes of machine-generated data (Log Files)

A log file

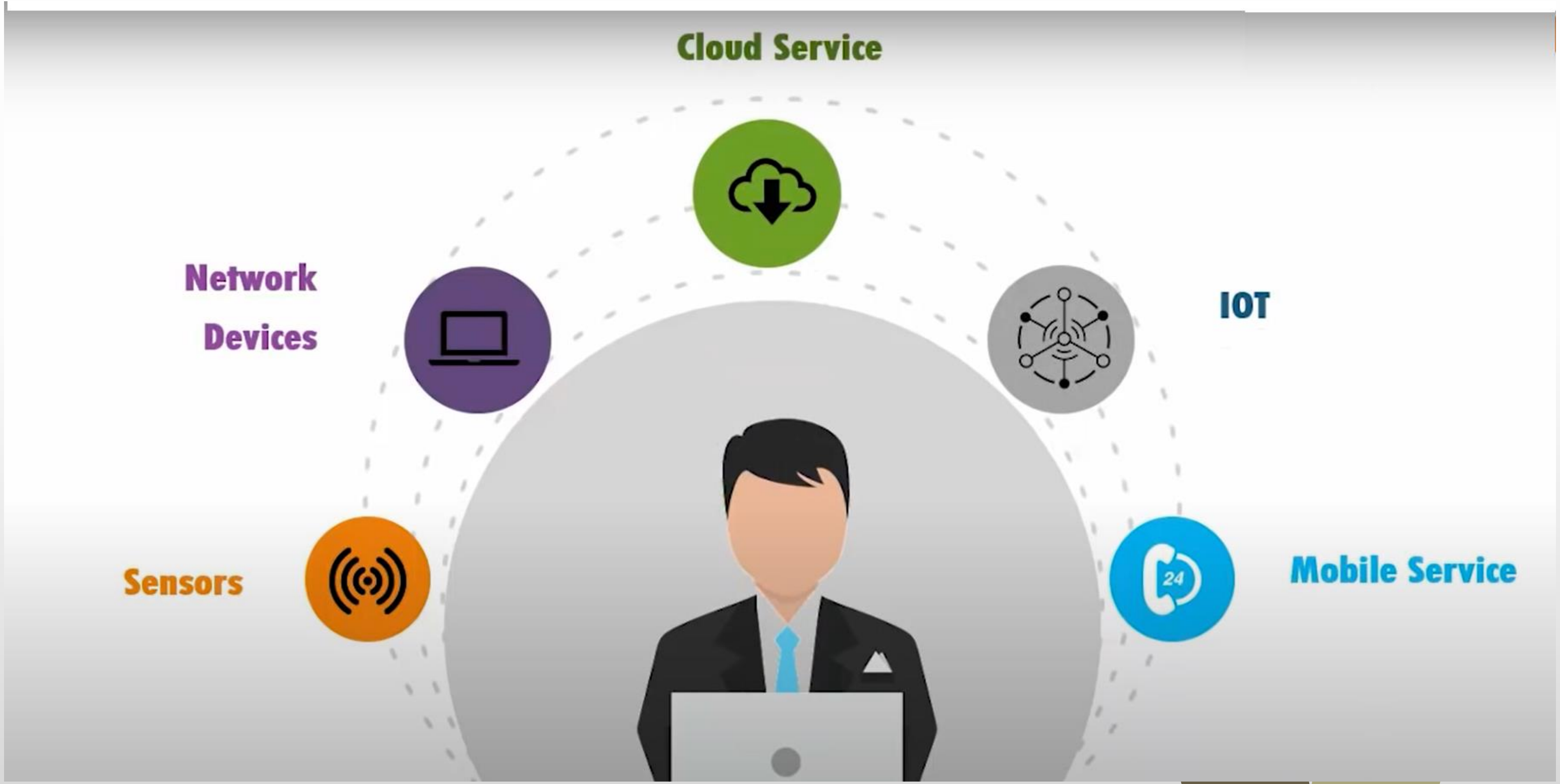
A log file (also known as machine data) is a computer-generated data file that contains information about usage patterns, activities, and operations within an **operating system**, **application**, **server** or **another device.**

log files are found

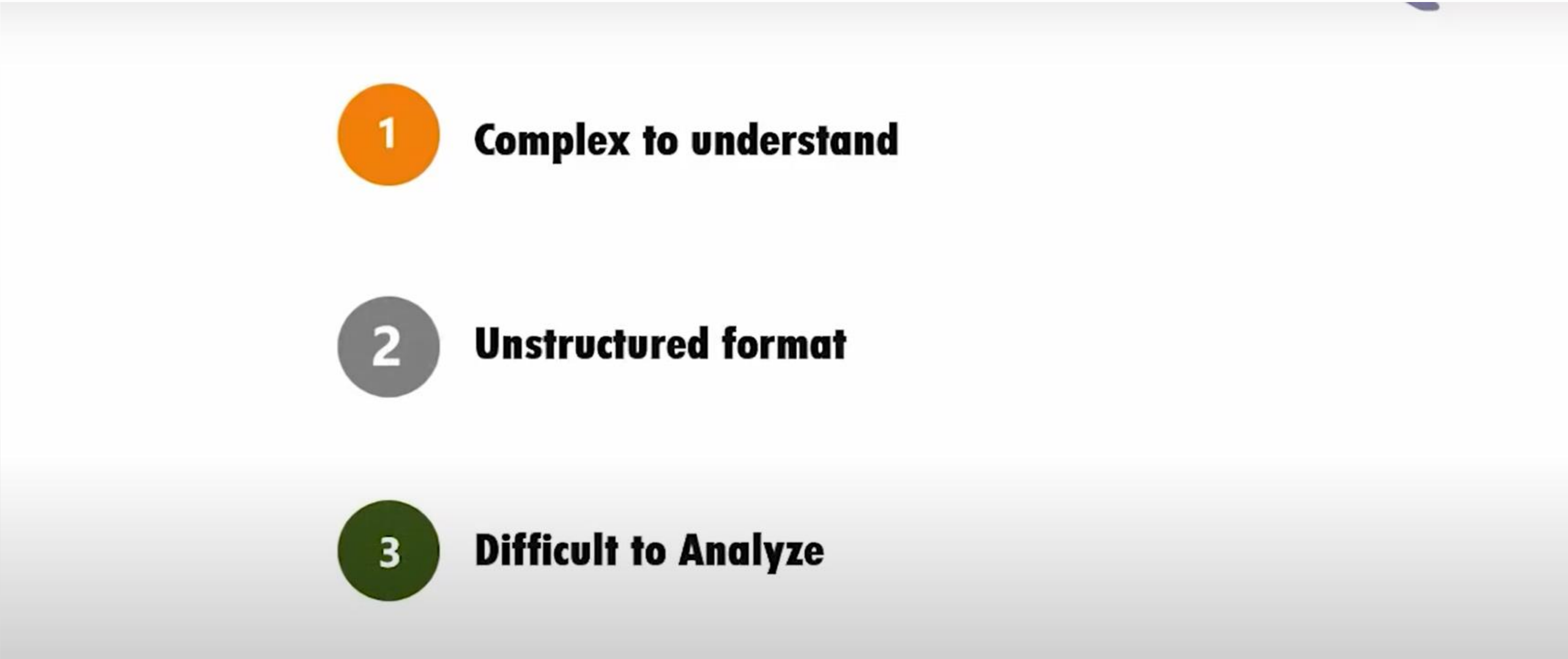
Operating systems
in applications,
web browsers,
hardware, and
even email etc.

Web Server Logs
Network Logs
Application Logs
Container Logs
System Logs
Security Logs

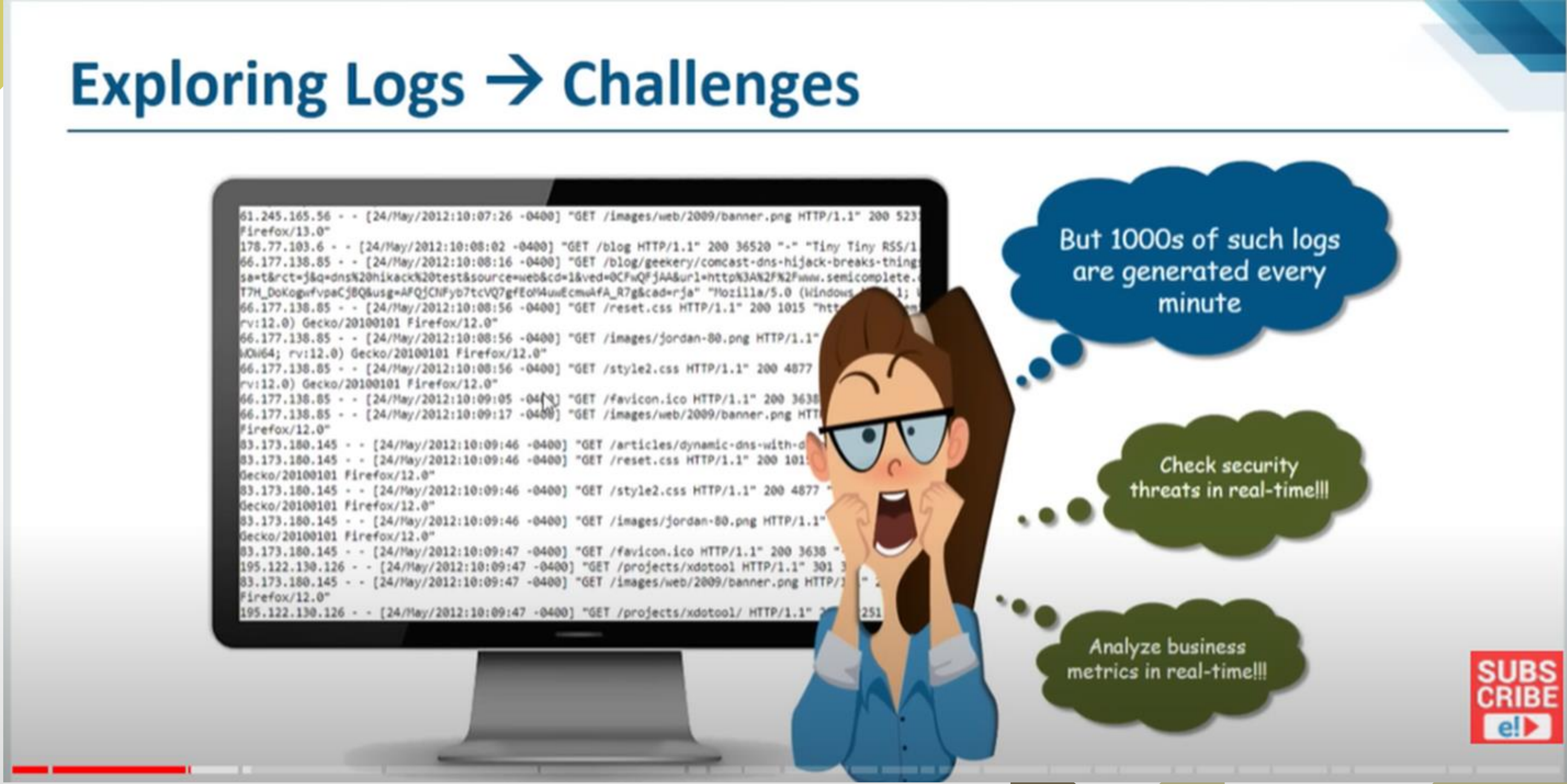
Log file from different sources



Challenge



1. Complex to understand
2. Unstructured format
3. Difficult to Analyze



Exploring Logs → Challenges

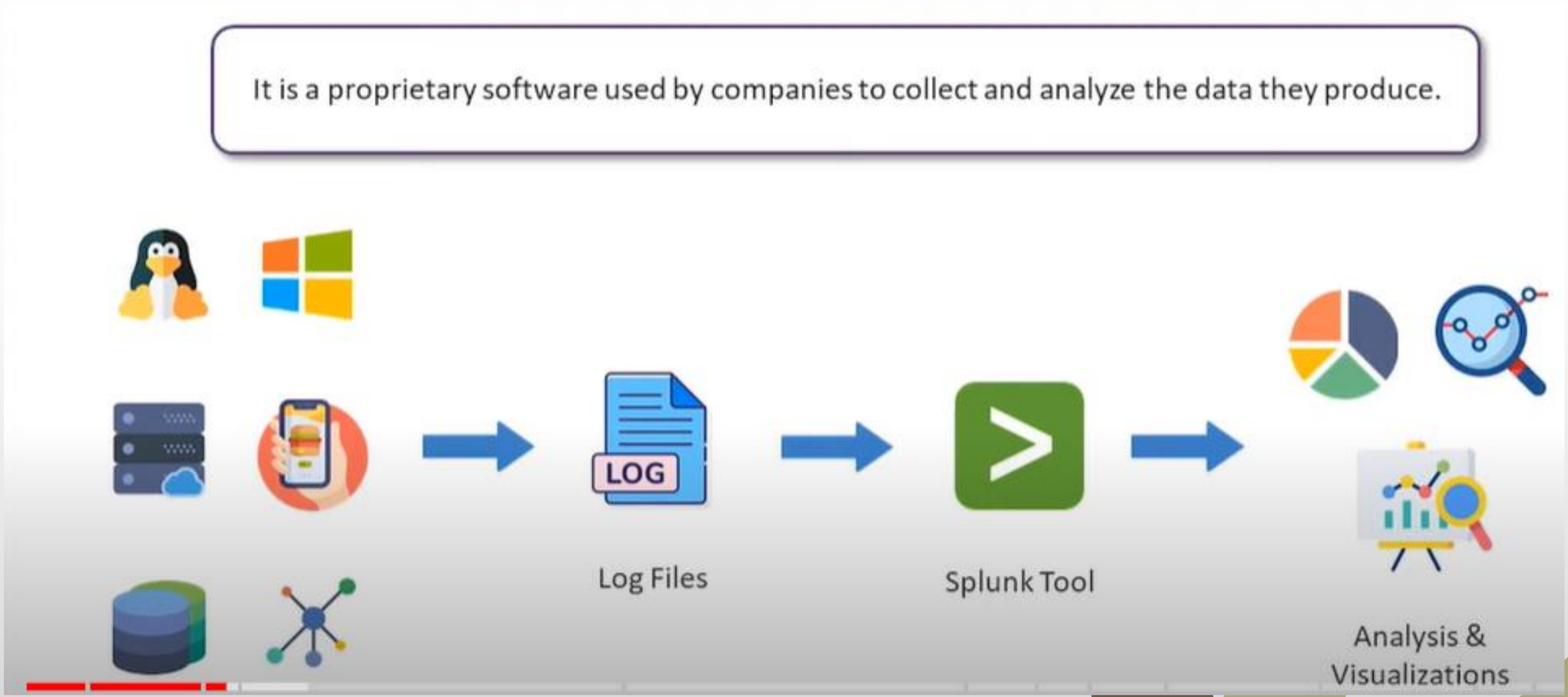
```
61.245.165.56 - - [24/May/2012:10:07:26 -0400] "GET /images/web/2009/banner.png HTTP/1.1" 200 523
Firefox/13.0"
178.77.103.6 - - [24/May/2012:10:08:02 -0400] "GET /blog HTTP/1.1" 200 36520 "-" "Tiny Tiny RSS/1
66.177.138.85 - - [24/May/2012:10:08:16 -0400] "GET /blog/geekery/comcast-dns-hijack-breaks-thing
sa=t&rct=j&q=dns%20hikack%20test&source=web&cd=1&ved=0CFwQFjAA&url=http%3A%2F%2Fwww.semicomplete.
T7H_DoKogwfvpaCjBQ&usg=AFQjCNFyb7tcVQ7gfEoM4uwEcmwAfA_R7g&cad=rja" "Mozilla/5.0 (Windows
66.177.138.85 - - [24/May/2012:10:08:56 -0400] "GET /reset.css HTTP/1.1" 200 1015 "htt
rv:12.0) Gecko/20100101 Firefox/12.0"
66.177.138.85 - - [24/May/2012:10:08:56 -0400] "GET /images/jordan-80.png HTTP/1.1"
WOW64; rv:12.0) Gecko/20100101 Firefox/12.0"
66.177.138.85 - - [24/May/2012:10:08:56 -0400] "GET /style2.css HTTP/1.1" 200 4877
rv:12.0) Gecko/20100101 Firefox/12.0"
66.177.138.85 - - [24/May/2012:10:09:05 -0400] "GET /favicon.ico HTTP/1.1" 200 3638
66.177.138.85 - - [24/May/2012:10:09:17 -0400] "GET /images/web/2009/banner.png HTT
Firefox/12.0"
83.173.180.145 - - [24/May/2012:10:09:46 -0400] "GET /articles/dynamic-dns-with-d
83.173.180.145 - - [24/May/2012:10:09:46 -0400] "GET /reset.css HTTP/1.1" 200 101
Gecko/20100101 Firefox/12.0"
83.173.180.145 - - [24/May/2012:10:09:46 -0400] "GET /style2.css HTTP/1.1" 200 4877 "
Gecko/20100101 Firefox/12.0"
83.173.180.145 - - [24/May/2012:10:09:46 -0400] "GET /images/jordan-80.png HTTP/1.1"
Gecko/20100101 Firefox/12.0"
83.173.180.145 - - [24/May/2012:10:09:47 -0400] "GET /favicon.ico HTTP/1.1" 200 3638 "
195.122.130.126 - - [24/May/2012:10:09:47 -0400] "GET /projects/xdotool HTTP/1.1" 301 3
83.173.180.145 - - [24/May/2012:10:09:47 -0400] "GET /images/web/2009/banner.png HTTP/1
Firefox/12.0"
195.122.130.126 - - [24/May/2012:10:09:47 -0400] "GET /projects/xdotool/ HTTP/1.1"
```

But 1000s of such logs are generated every minute

Check security threats in real-time!!!

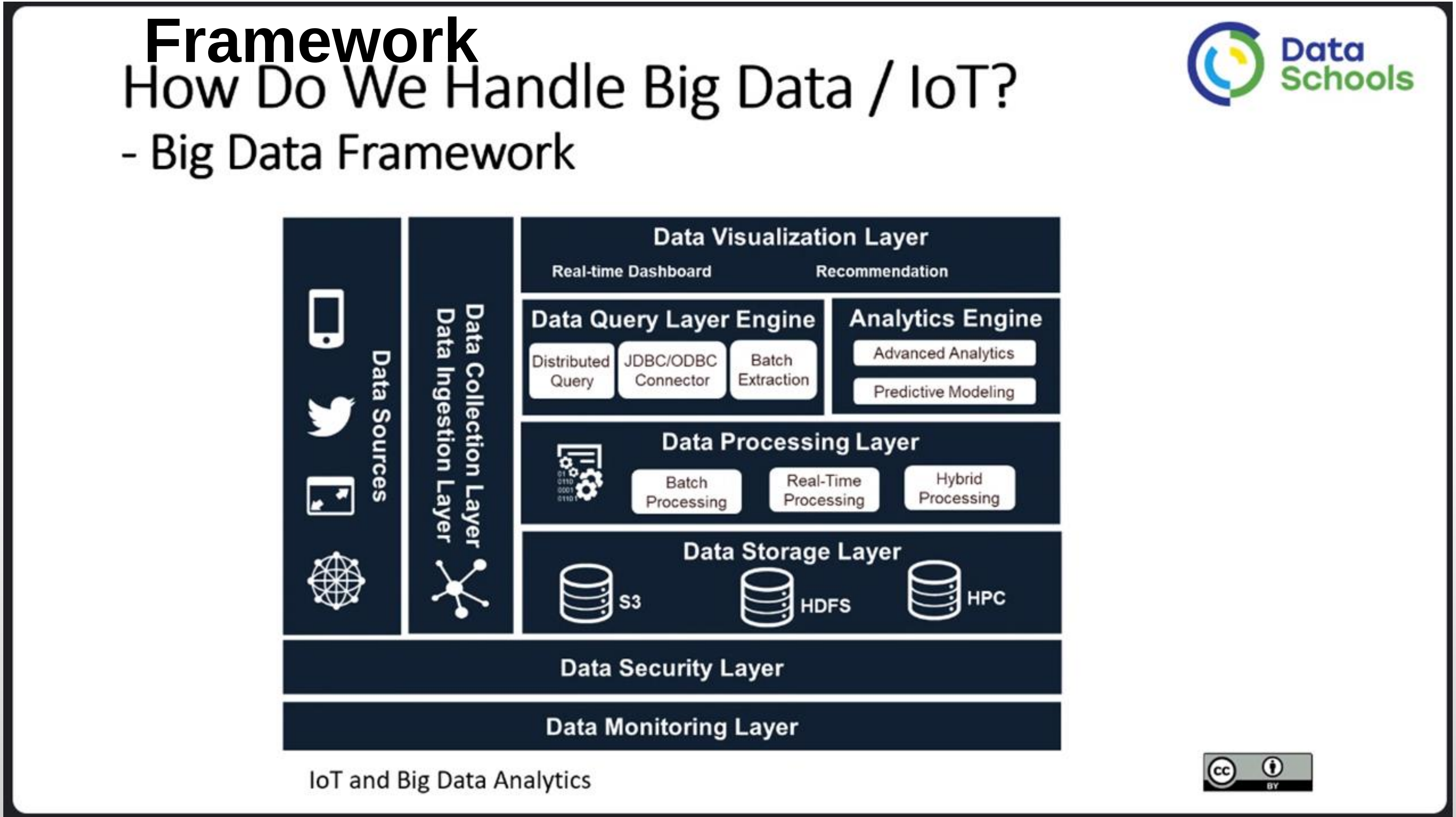
Analyze business metrics in real-time!!!

Splunk for log file



It is a proprietary software used by companies to collect and analyze the data they produce.

Seeing the bigger picture: BIG Data Framework



How Do We Handle Big Data / IoT?

- Big Data Framework

Data Sources

Data Collection Layer
Data Ingestion Layer

Data Visualization Layer
Real-time Dashboard
Recommendation

Data Query Layer Engine
Distributed Query
JDBC/ODBC Connector
Batch Extraction

Analytics Engine
Advanced Analytics
Predictive Modeling

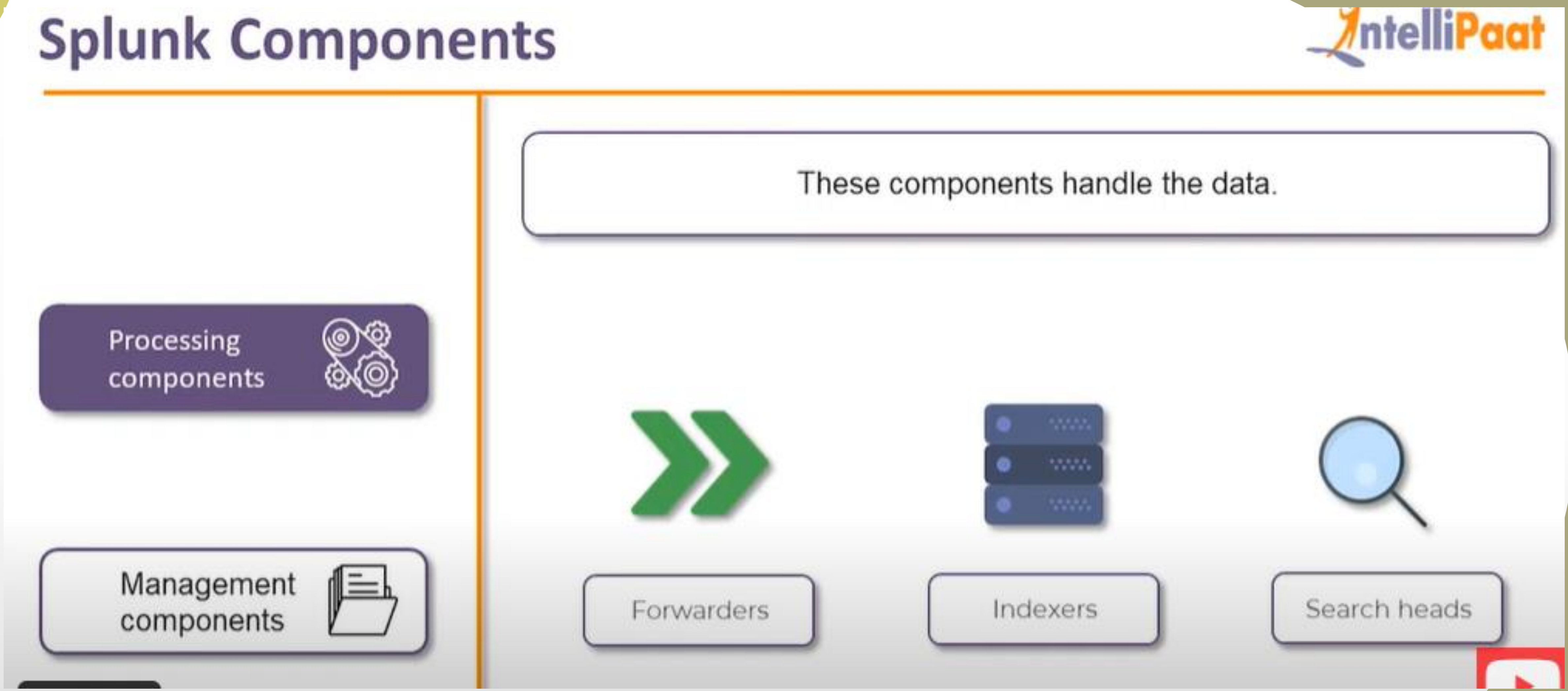
Data Processing Layer
Batch Processing
Real-Time Processing
Hybrid Processing

Data Storage Layer
S3
HDFS
HPC

Data Security Layer

Data Monitoring Layer

IoT and Big Data Analytics



Splunk Components

These components handle the data.

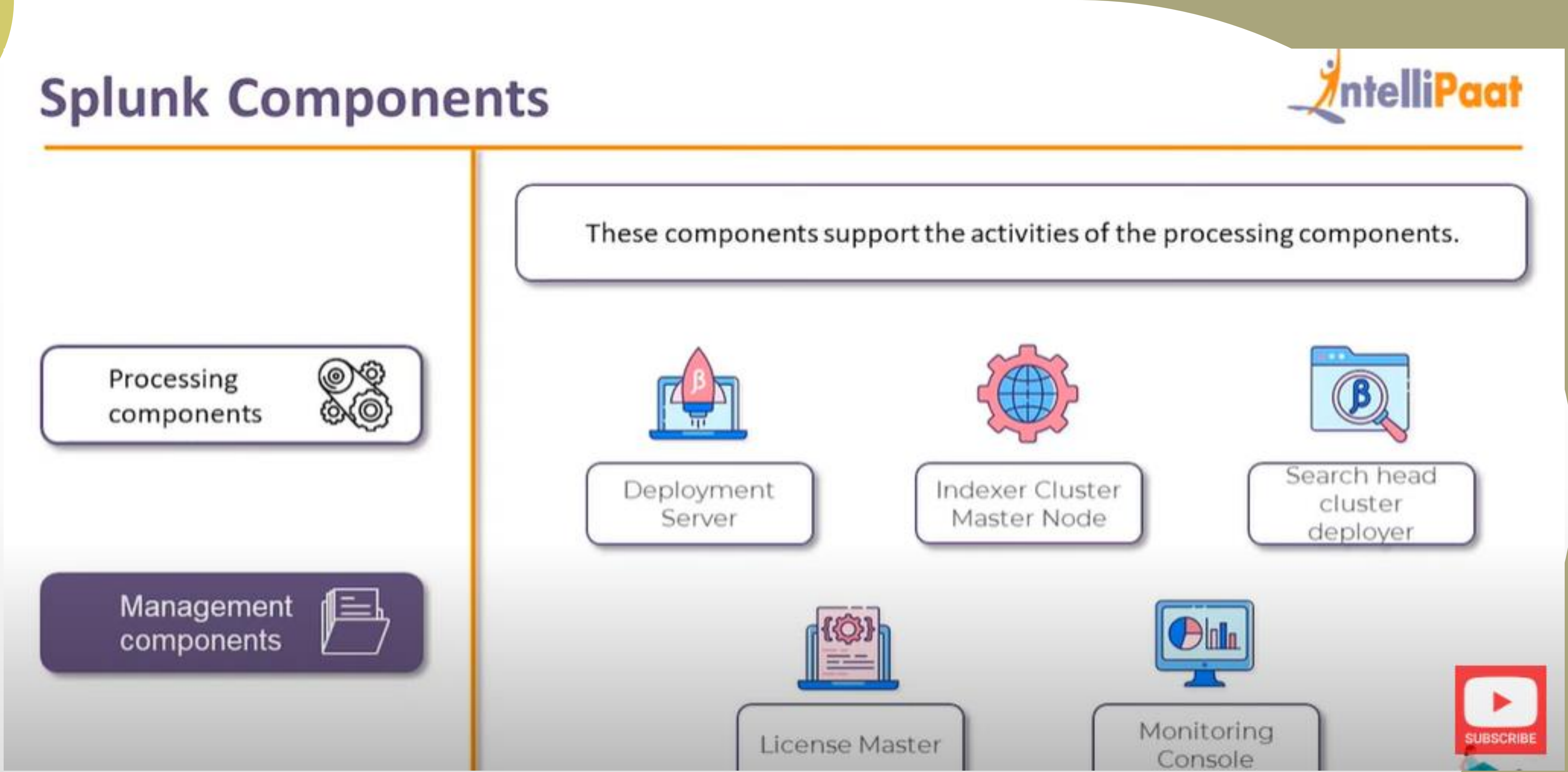
Processing components

Management components

Forwarders

Indexers

Search heads



Splunk Components

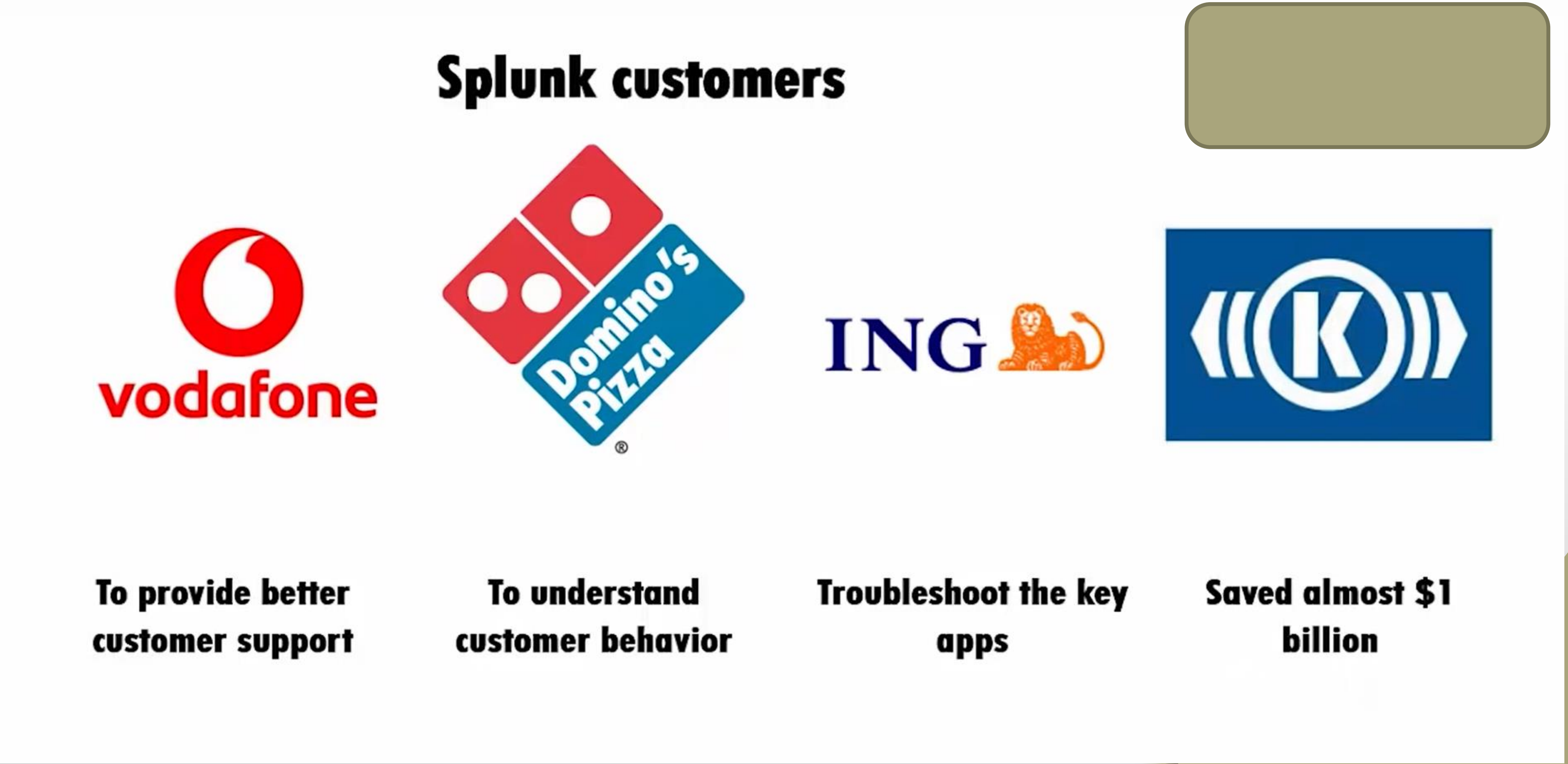
Processing components

Management components

These components support the activities of the processing components.

- Deployment Server
- Indexer Cluster Master Node
- Search head cluster deployer
- License Master
- Monitoring Console

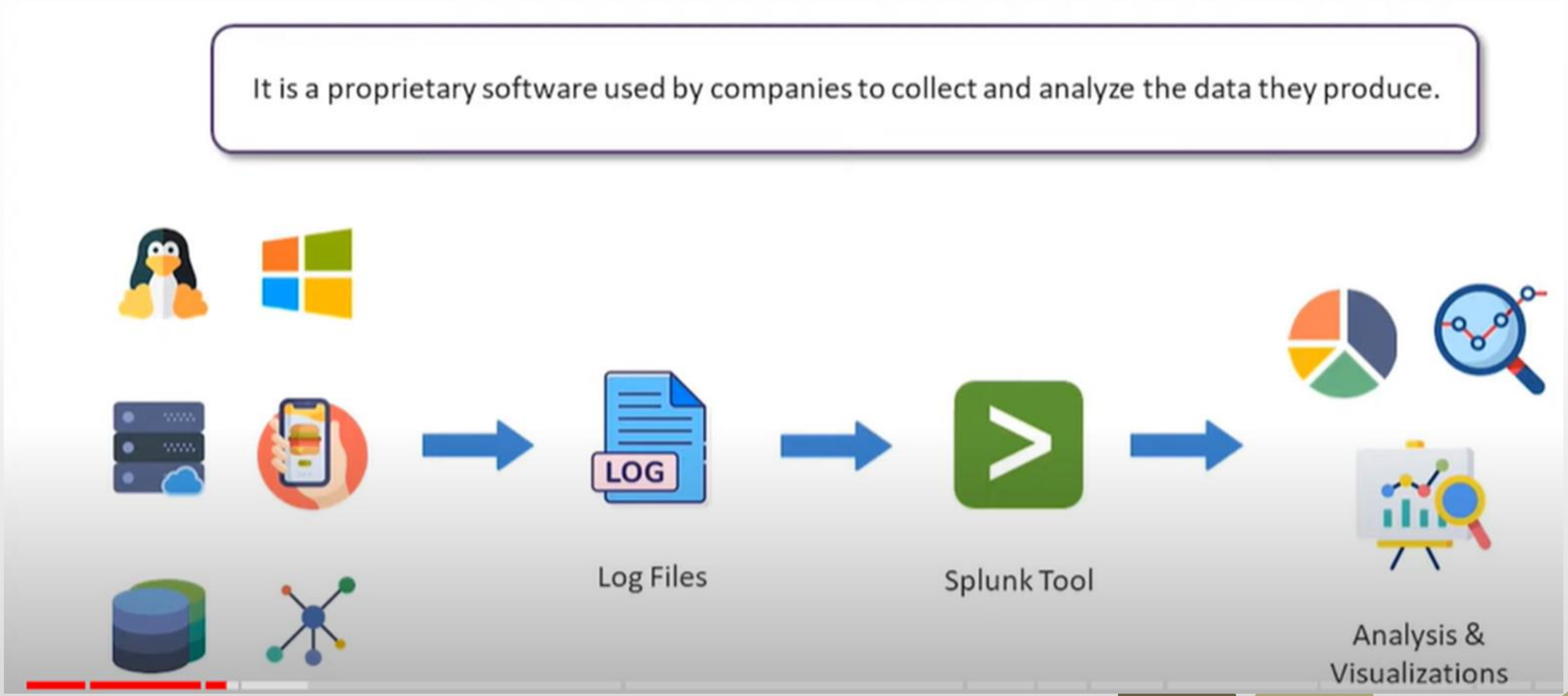
Some Splunk Customers


vodafone
Domino's Pizza
ING

Splunk customers

To provide better customer support	To understand customer behavior	Troubleshoot the key apps	Saved almost $1 billion

Analysis and Visualization of Log Data



It is a proprietary software used by companies to collect and analyze the data they produce.

Thank you for your Attentions!