IOT-BASED PLATFORM
AIR QUALITY (PM2.5) MONITORING SYSTEM

Big Data Tool with Apache Kylin

Project Members:

Suyogya Ratna Tamrakar (st121334)
Younten Tshering (st121775)
Smrity Baral (st121662)
Shubhangini Gontia (st121473)

Outlines

1. Overview on Big Data
2. Overview on Hadoop
3. Apache Kylin
4. Updated System Architecture
5. Process of getting started with Kylin
6. Update on the project
7. Subsystem decomposition
8. Access Matrix
9. Future Work

Overview on Big Data

Now, before moving on to **Apache Kylin**, let us start the with **Big Data**, that led to the development of **Hadoop** and **then** to Apache Kylin.

Big Data is a problem statement

1. The first problem is storing the huge amount of data.

Storing huge data in a traditional system is not possible. The reason is obvious, the storage will be **limited to one system** and the data is increasing at a **tremendous rate.**

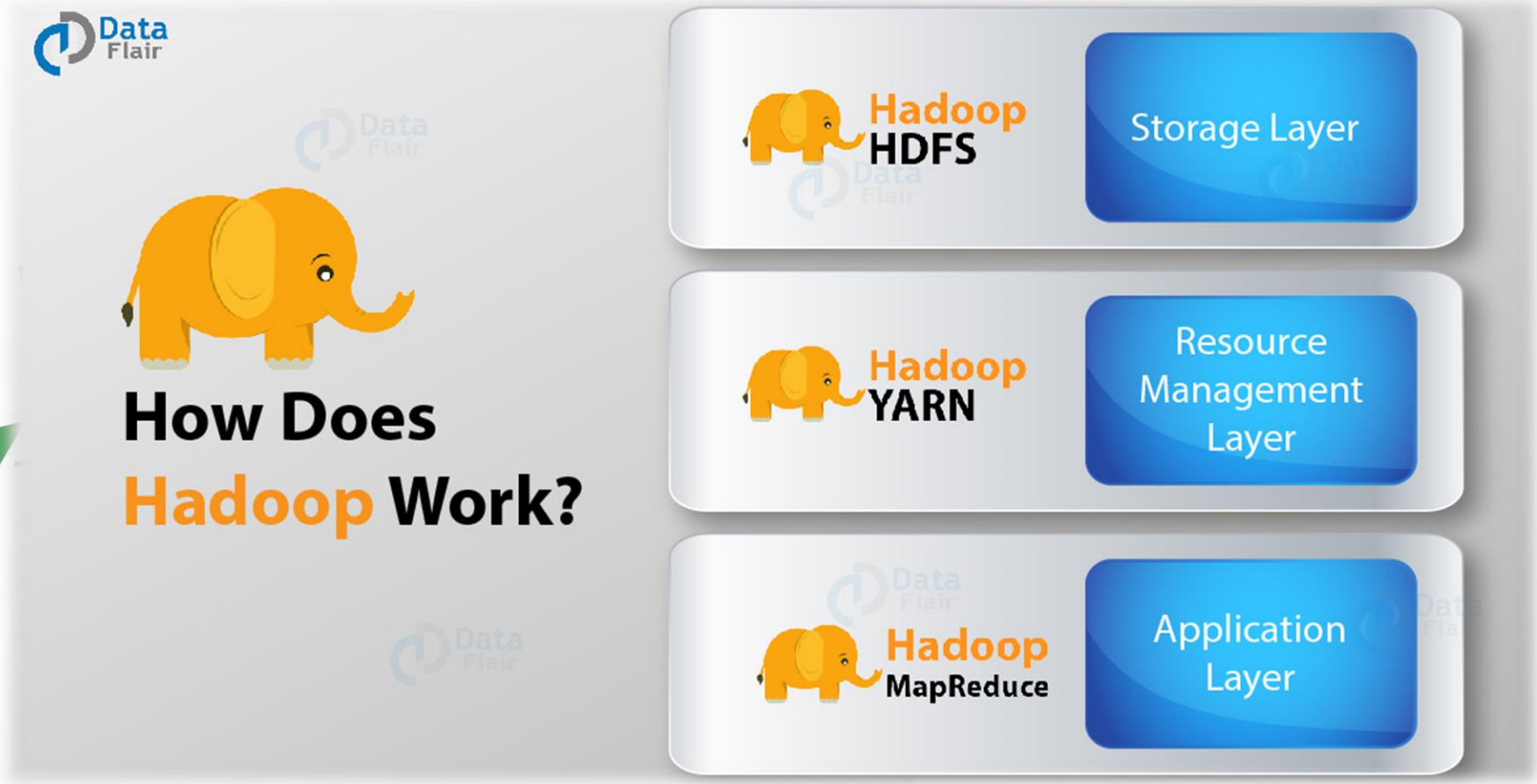
2. The second problem is storing heterogeneous data.

The data is not only huge, but it is also present in **various formats i.e. unstructured, semi-structured and structured**. So, we need to make sure that we have a system to store different types of data that is generated from various sources.

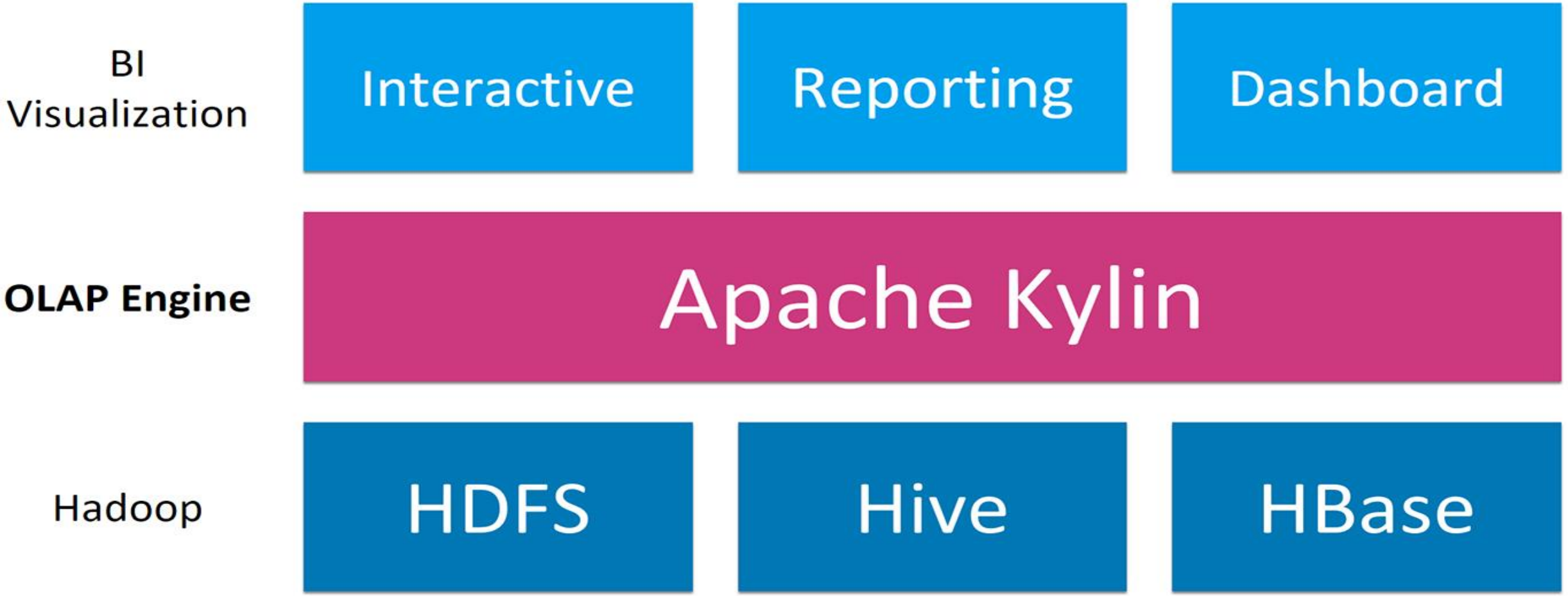
3. Finally, the third problem, which is the processing speed.

Now the **time taken to process** this huge amount of data is quite high as the data to be processed is too large.

Overview on Hadoop



Hadoop Eco-system

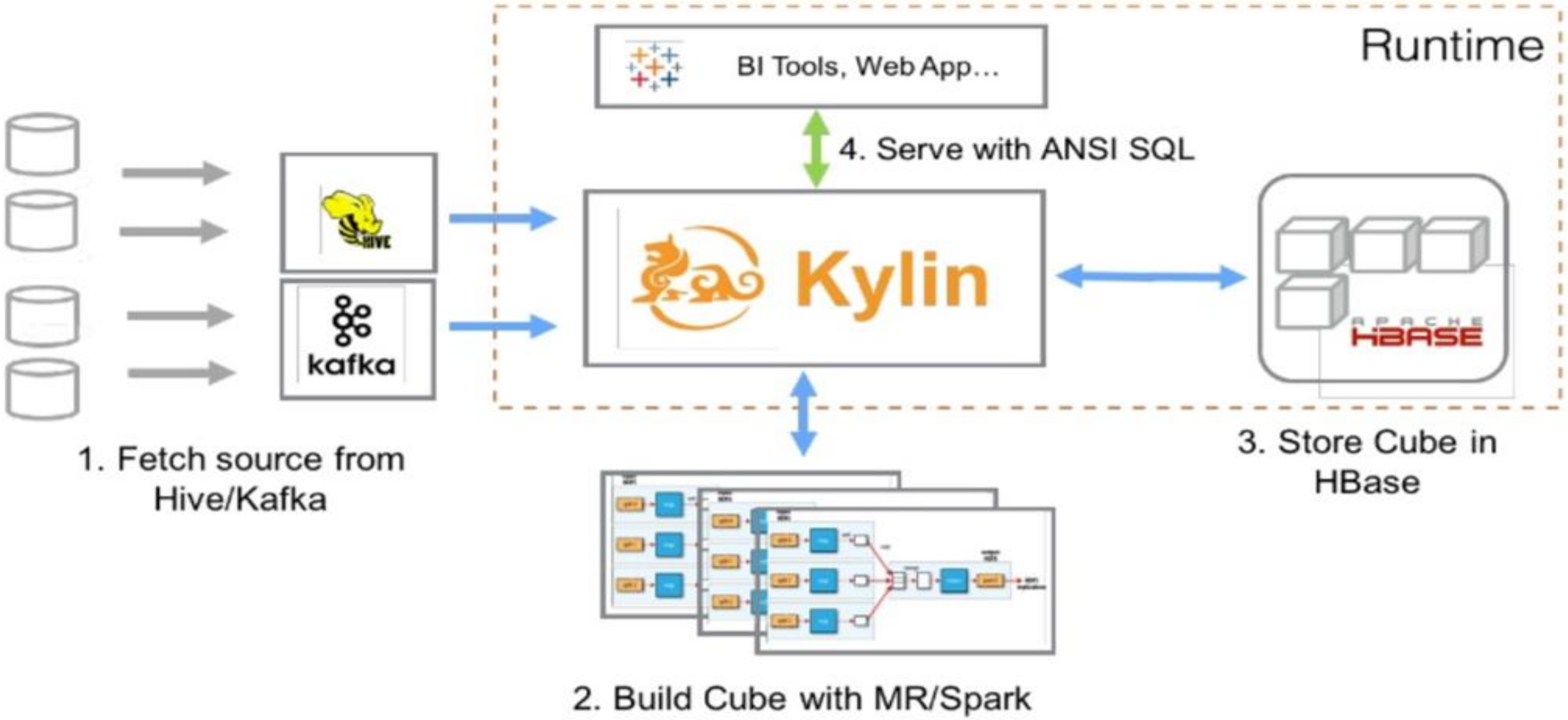


Apache Kylin

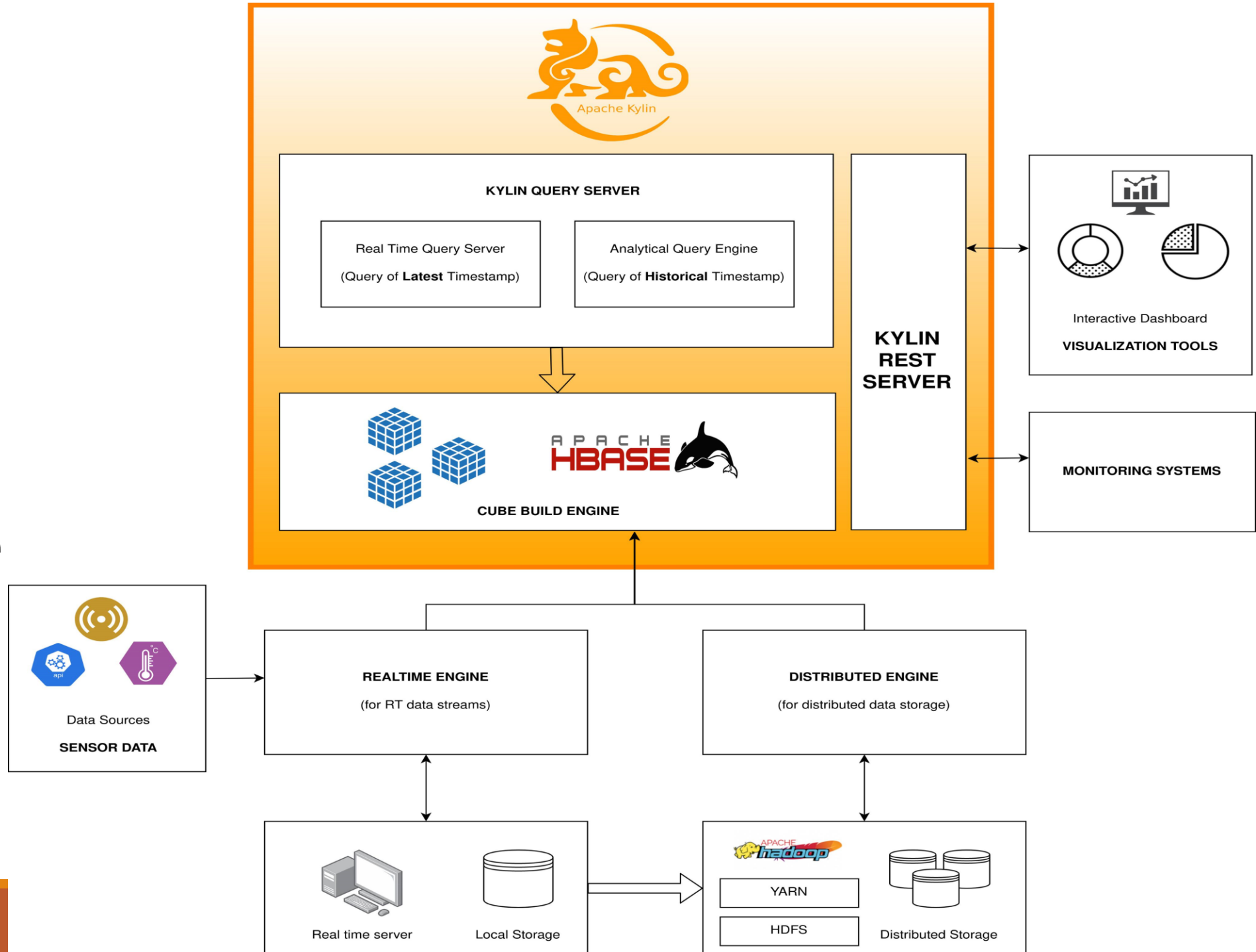
OLAP on Hadoop – Apache Kylin

Kylin is an **open source Distributed Analytical Engine** that provides SQL interface and multidimensional analysis (OLAP) on Hadoop supporting extremely large datasets. Apache kylin pre-calculates OLAP cubes and store the cubes into a reliable and scalable datastore (**HBase**).

Apache Kylin: How it works?

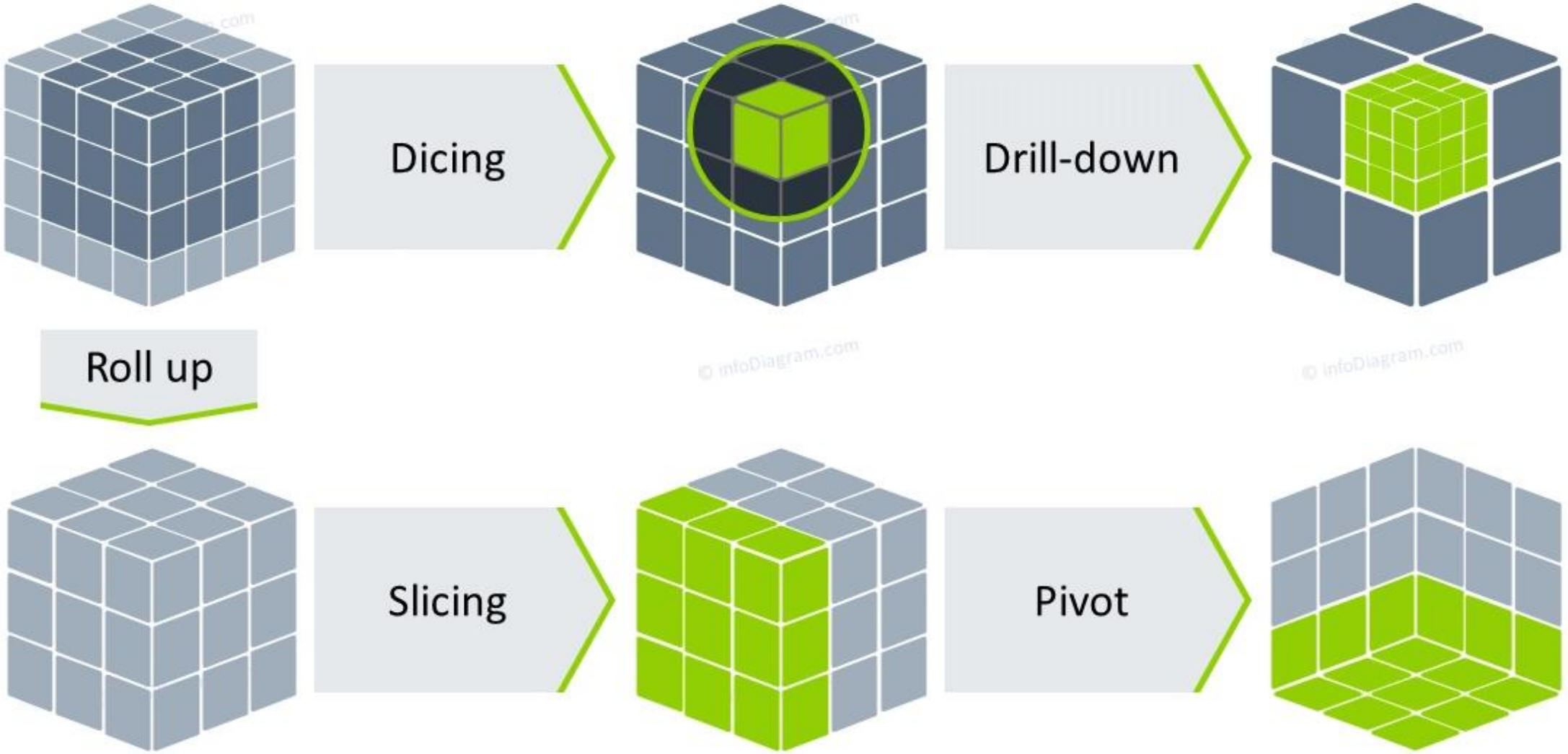


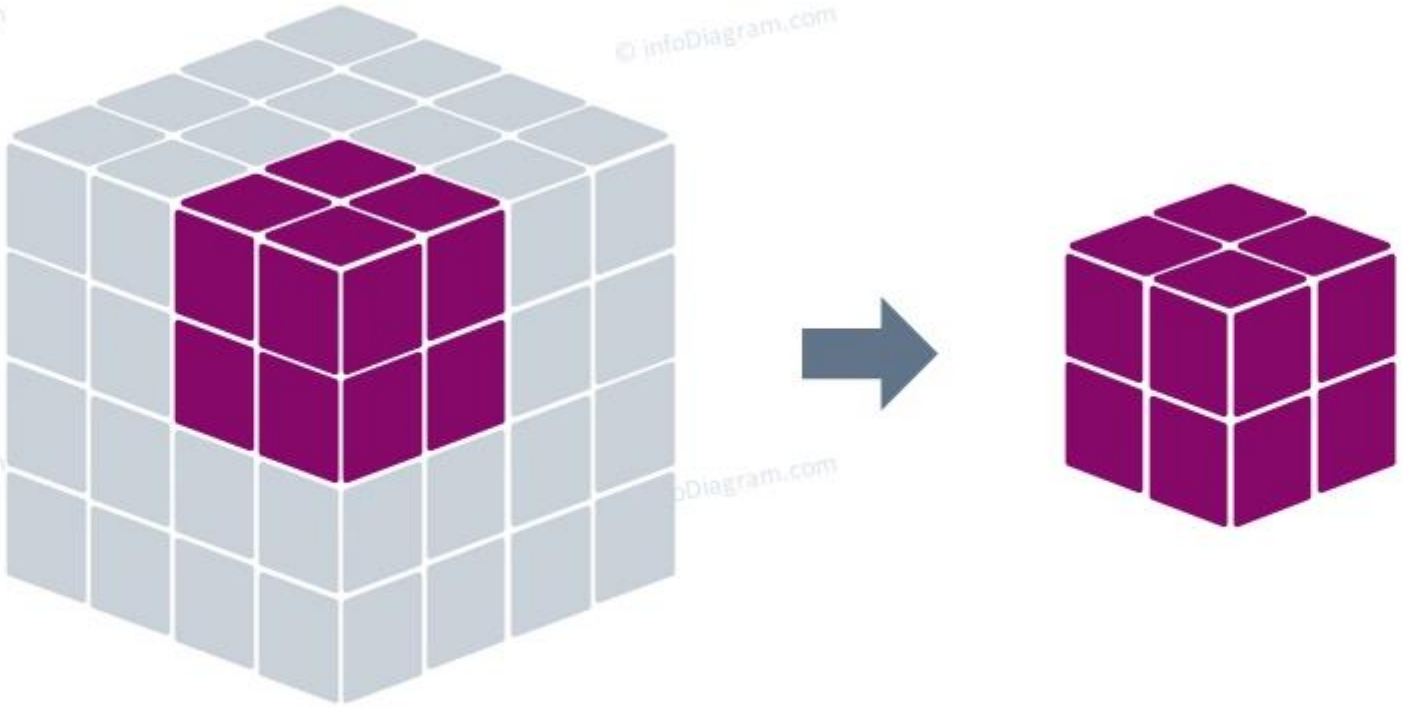
Updated System Architecture



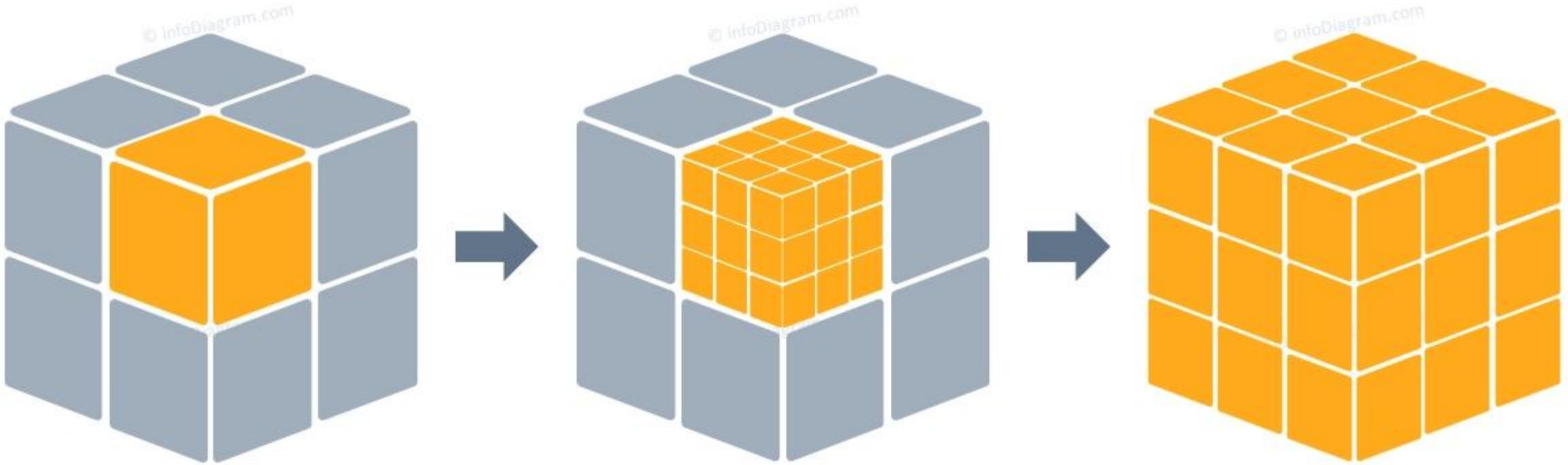
Operations on OLAP Cube

Illustrations of Dicing, Drill-down, Roll-up, Slicing, Pivot

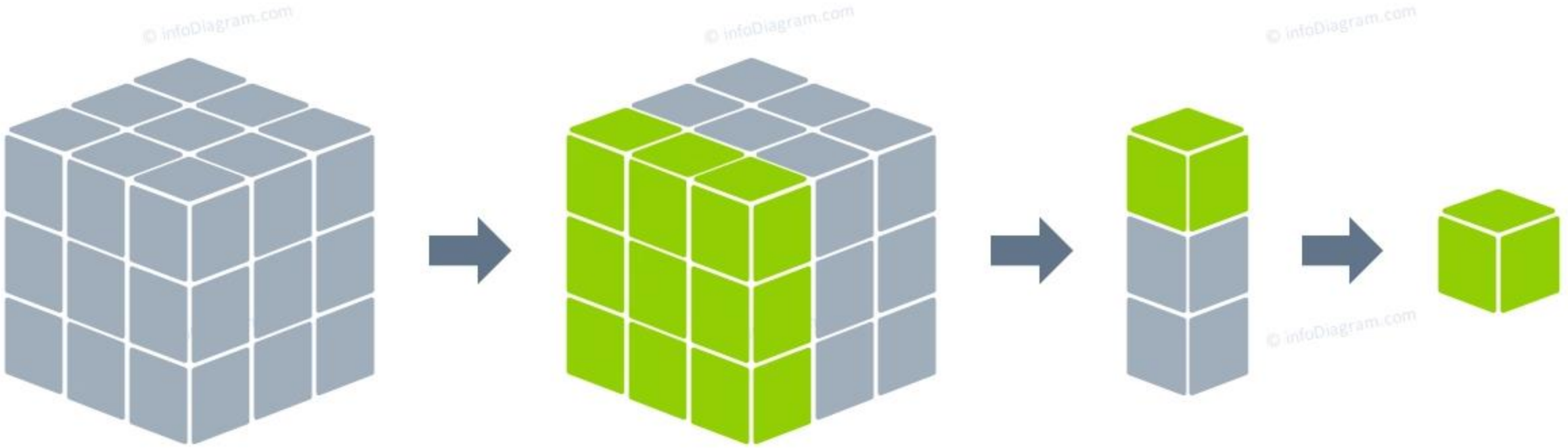




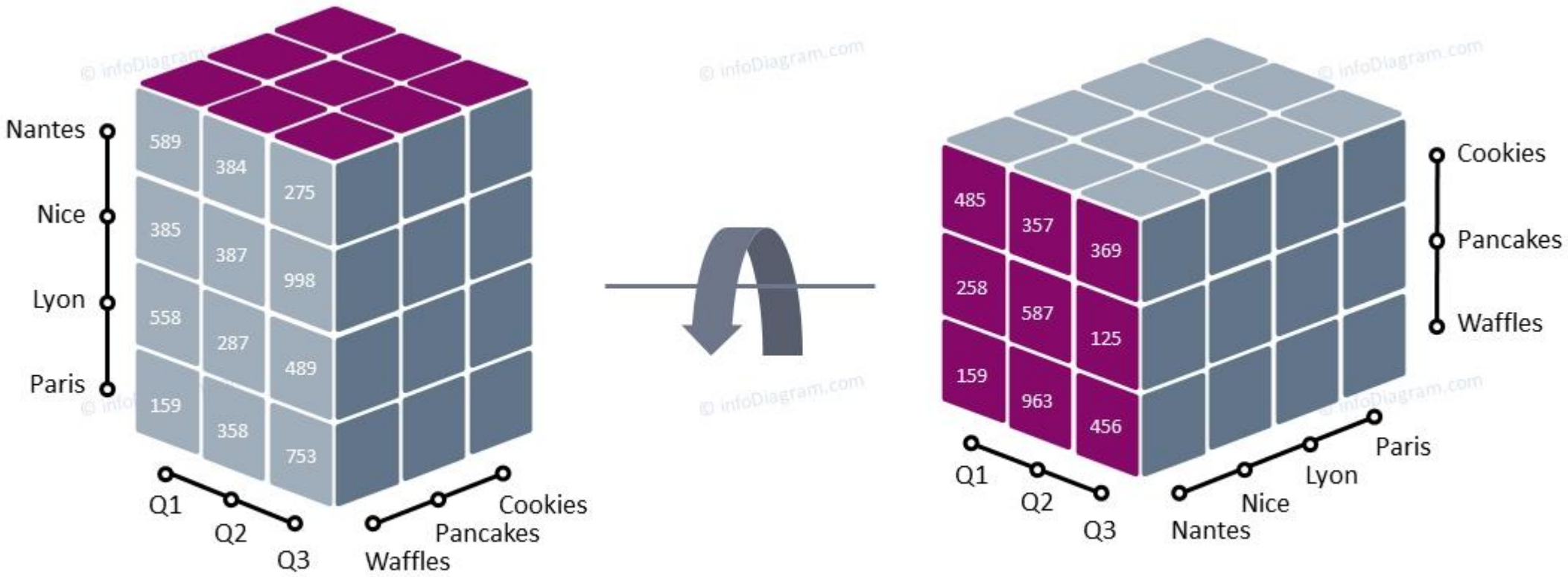
Dicing is an operation of creating a sub-cube from the main one.



Drill-down is an operation opposite to roll-up.



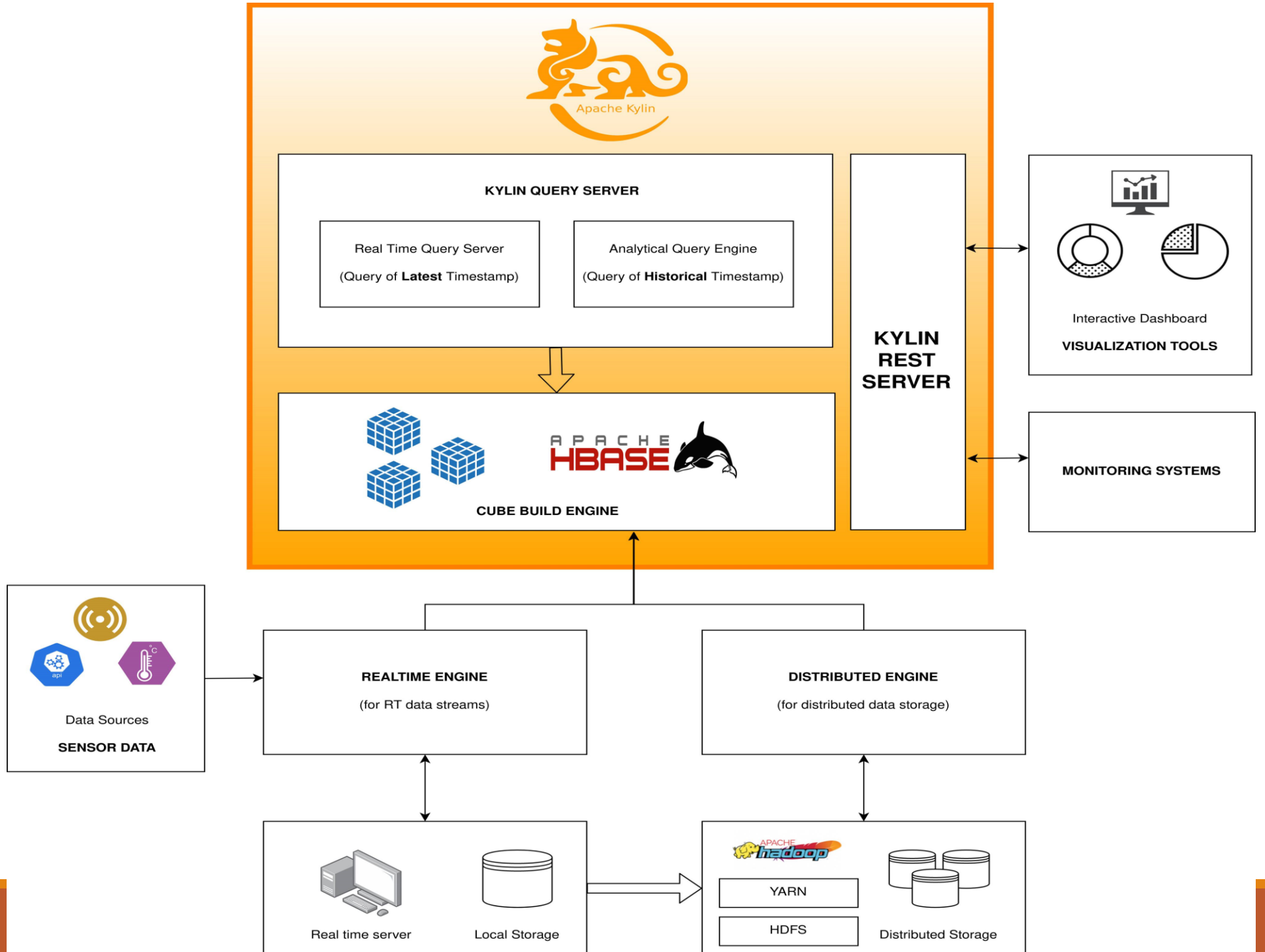
Slicing is subtracting rectangular part of a cube of the same single value



Pivoting allows to see another perspective on the dataset, by rotating the whole cube in space.

Roll-up - OLAP Operation Explanation Template

Summarizing data along a dimension (unlike dicing, it's not picking a sub-cube).


Apache Kylin
KYLIN QUERY SERVER
Real Time Query Server
(Query of Latest Timestamp)
Analytical Query Engine
(Query of Historical Timestamp)
CUBE BUILD ENGINE
KYLIN REST SERVER
Interactive Dashboard
VISUALIZATION TOOLS
MONITORING SYSTEMS
Data Sources
SENSOR DATA
REALTIME ENGINE
(for RT data streams)
DISTRIBUTED ENGINE
(for distributed data storage)
Real time server
Local Storage
YARN
HDFS
Distributed Storage

Process of getting started with Kylin

Kylin Installation:

Software Requirements

- Hadoop: 2.7+, 3.1+ (since v2.5)
- Hive: 0.13 - 1.2.1+
- HBase: 1.1+, 2.0 (since v2.5)
- Spark (optional) 2.3.0+
- Kafka (optional) 1.0.0+ (since v2.5)
- JDK: 1.8+ (since v2.5)
- OS: Linux only, CentOS 6.5+ or Ubuntu 16.0.4+

- Hadoop

Hadoop is an open-source software framework for storing data and running applications on clusters of commodity hardware.

- Hive

Apache Hive is an open source data warehouse software for reading, writing and managing large data set files that are stored directly in either the Apache Hadoop Distributed File System (HDFS)

- HBase

Apache HBase is used to have random, real-time read/write access to Big Data. It hosts very large tables on top of clusters of commodity hardware

- Kafka

Kafka is used for real-time streams of data, to collect big data, or to do real time analysis (or both).

❖ JDK and Linux

Hardware Requirement

- The minimum configuration of a server running Kylin is 4 core CPU, 16 GB RAM and 100 GB disk. For high-load scenarios, a 24-core CPU, 64 GB RAM or higher is recommended.

- Kylin relies on Hadoop clusters to handle large data sets. You need to prepare a Hadoop cluster with HDFS, YARN, MapReduce, Hive, HBase, Zookeeper and other services for Kylin to run.

Kylin can be launched on any node in a Hadoop cluster.

- Download Kylin from below link and
- https://kylin.apache.org/download/
- Once installed, start kylin by following command
- $KYLIN_HOME/bin/kylin.sh start
- You can open the GUI from
- http://<hostname>:7070/kylin

Kylin Image from Docker hub

- Docker is a set of platform as a service products that use OS-level virtualization to deliver software in packages called containers. Containers are isolated from one another and bundle their own software, libraries and configuration files

➢ docker pull apachekylin/apache-kylin-standalone:3.1.0

Update on the project

github.com/shubhanginigon/SDQI2021_G1

Gmail YouTube Maps AIT site Images %Temp% Important Helpful site

main 4 branches 0 tags Go to file Add file Code

YountenTshering Merge pull request #15 from shubhanginigon/YountenTshering ... df45da3 2 hours ago 37 commits

.idea	Added sidebar navigation	8 days ago
.vscode	created packages and user.java	10 days ago
AQMS	Merge pull request #13 from shubhanginigon/YountenTshering	6 days ago
Project Kick-off Presentation File	Upload file and update readme file	12 days ago
Software Requirement Specification P...	Upload file and update readme file	12 days ago
Updated Document and Progress rep...	New Folder for progress and update readme file	2 hours ago
img	New Folder for progress and update readme file	2 hours ago
README.md	New Folder for progress report and updated readme	2 hours ago

README.md

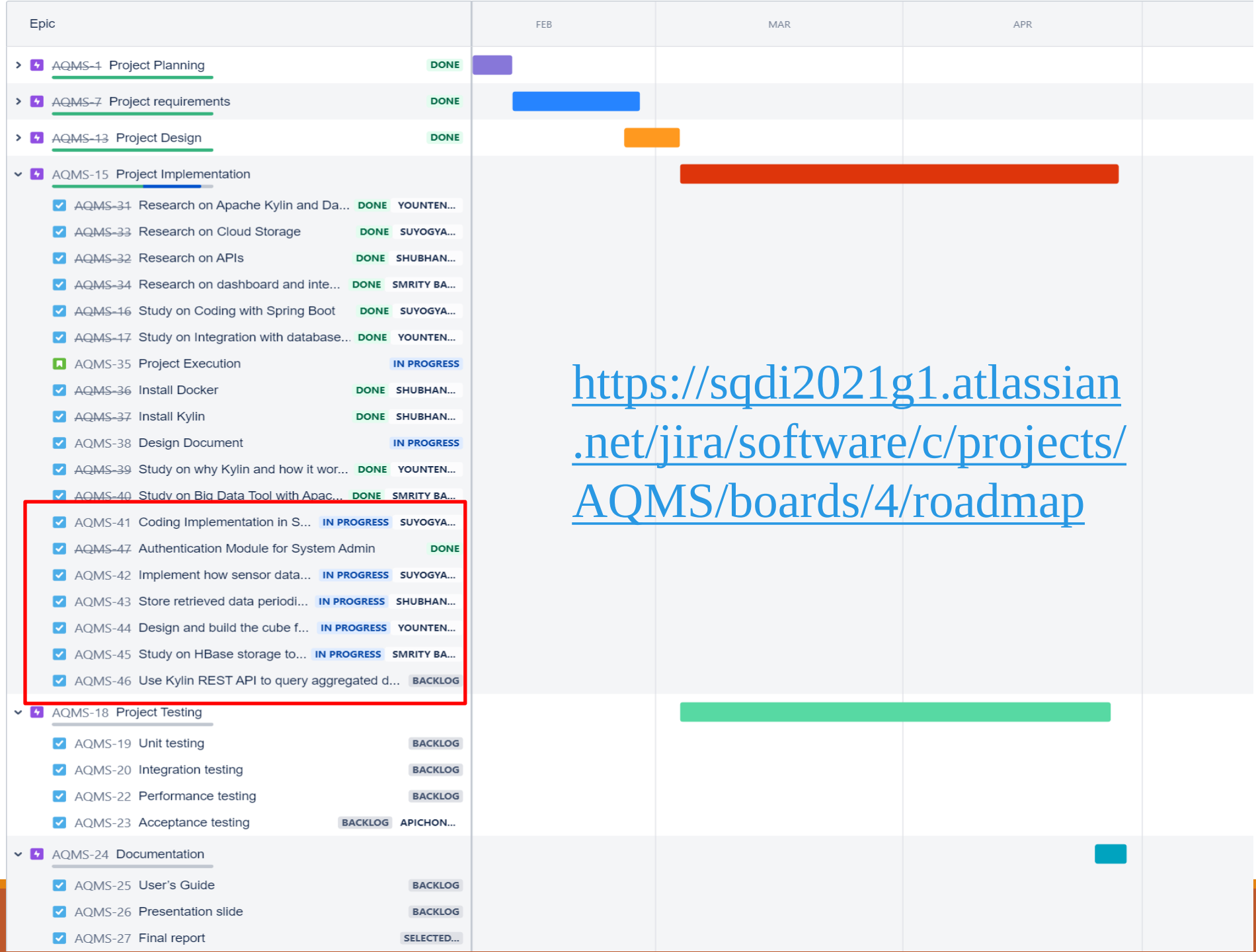
GROUP 1 : IOT BASED PLATFORM

Group Members:

1. Smrity Baral (st121662)
2. Shubhangini Gontia (st121473)
3. Suyogya Ratna Tamrakar (st121334)
4. Younten Tshering (st121775)

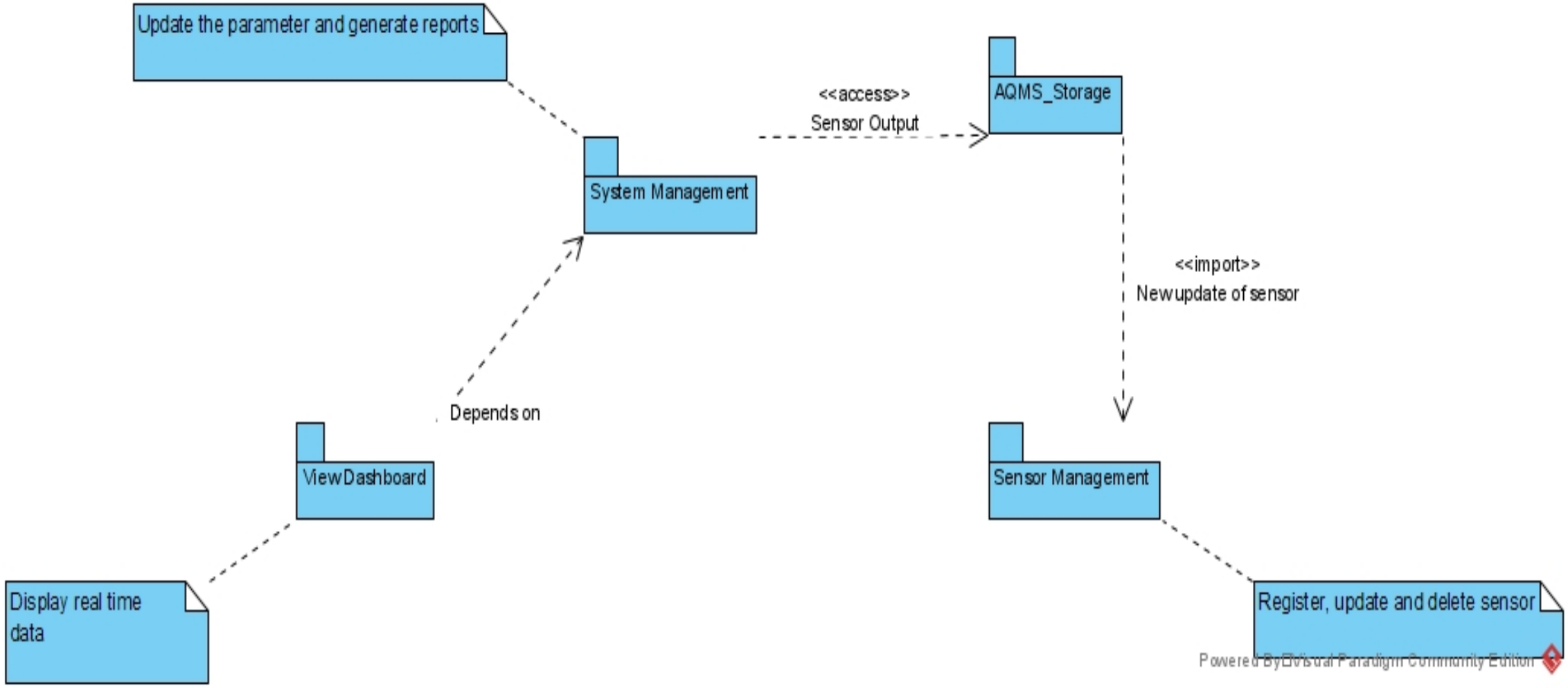
https://github.com/shubhanginigon/SDQI2021_G1

Schedule



https://sqdi2021g1.atlassian.net/jira/software/c/projects/AQMS/boards/4/roadmap

Subsystem decomposition



Access Matrix

Access Rights

Classes

Table 1. Access Matrix

Actors

	Dashboard	**System Setting**
Admin	viewDashboard () generateReport ()	registerSensor () updateParameter () activateSensor ()
Users (Public)	viewDashboard ()	

Future Work

- Integration with Hadoop
- Some part of Design Document
- Project Testing
- Final Project Documentation

References

Apache Kylin. (2015). Bring OLAP back to big data! Retrieved from Apache Kylin | Analytical Data Warehouse for Big Data

Fann,N.,& Risley,D. (2011,January 5). The public health context for PM2.5 and ozone air quality trends. Air Qual Atmos Health 6, 1–11 (2013). https://doi.org/10.1007/s11869-010-0125-0

Geetha,S.M.N. (2021, March 19). Hadoop for Analyst-Apache Druid, Apache Kylin and Interactive query tools. Retrieved from https://www.saigeetha.in/post/hadoop-for-analysts-apache-druid-apache-kylin-and-interactive-query-tools?fbclid=IwAR0RRXXxKmv8onswnS-g5mV5Hh_L5R9zOSWly6YO8d4kb6oYYW4rrjF5wlo

Gupta,A.k., & Johari,R. (2019). IOT based electrical device surveillance and control system. International Conference on Internet of Things: Smart Innovation and Usages (IoT-SIU), https://doi.org/10.1109/IoT-SIU.2019.8777342

Nethu, M.V. (2018, September 25). OLAP on Hadoop-Apache Kylin. Retrieved from https://medium.com/@mvneethu90/olap-in-hadoop-apache-kylin-bf0377d8b44f

Sinha, S. (2016, October 28). Hadoop ecosystem- Get to know the Hadoop tools for crunching Big Data. edureka. Retrieved from https://medium.com/edureka/hadoop-ecosystem-2a5fb6740177

Sinha,S. (2014, October 9). Hadoop tutorial- A comprehensive guide to Hadoop. edureka. Retrieved from https://medium.com/edureka/hadoop-tutorial-24c48fbf62f6

Questions and Feedback