aws

COMMUNITY DAY

PUNE 2022

aws

COMMUNITY DAY

PUNE 2022

Building application with analytical capabilities using AWS services and Apache Druid

druid®

Apache®, Apache Druid®, Druid®, and the Druid logo are either registered trademarks or trademarks of The Apache Software Foundation in the United States and other countries.
This presentation is not affiliated with, endorsed by, or otherwise associated with the Apache Software Foundation (ASF) or any of its projects.

Before we start…

How many of you heard of Druid

How many of you are tried Druid in the past

How many of you are using / planning to use druid in production

How many of you have the use case of providing analytics tool to your users as part of the product offerings

aws
COMMUNITY DAY
PUNE 2022

Again we are not starting ..

How many of you are application developers

How many of you are Data Scientist

How many of you are Data Engineers

aws
COMMUNITY DAY
PUNE 2022

Who Am I

Tijo Thomas

Lead Solutions Architect @ Imply

Dealing with Big Data Challenges since 2014

tijopthomas

@tijothomas214 tijopthomas

aws
COMMUNITY DAY
PUNE 2022

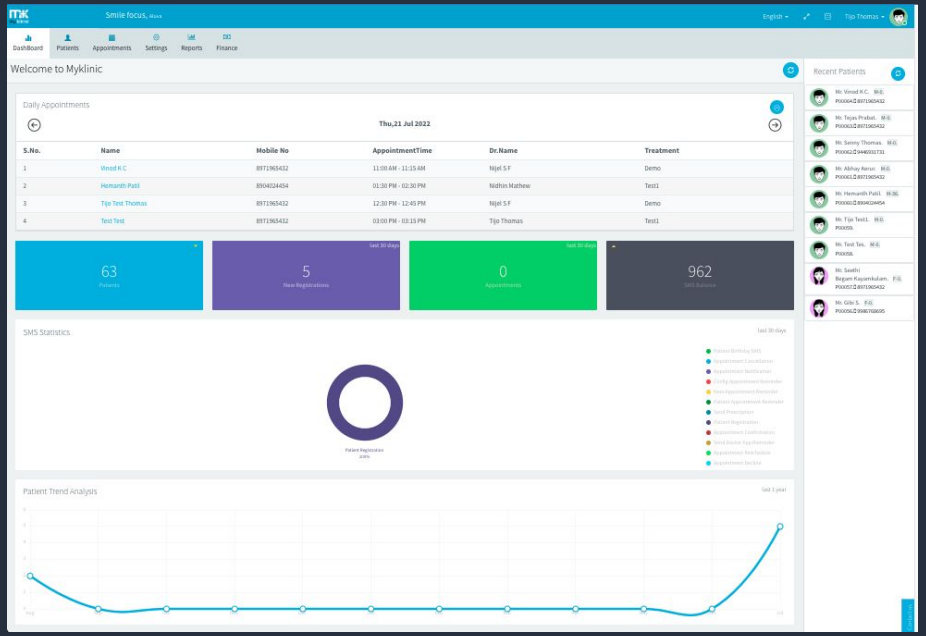
Agenda

- Story of building an analytical app
- Demo - Wiring it up with AWS Services
- Demo on performance and concurrency
- How is Druid is best fit analytical apps
- Q&A

I want to build…

“ In my application landing page, I want to show a real-time dashboard”

I want to build dashboard



Pain Areas

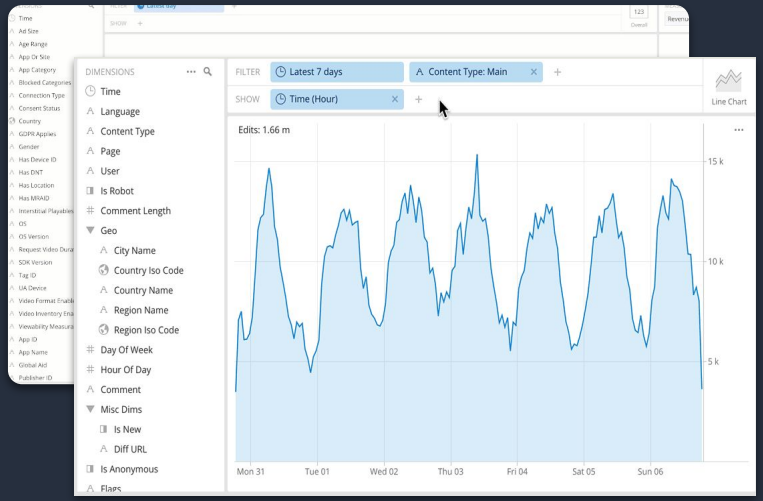
- This dashboard aggregate data over wide range of data points
- Dashboard refreshes frequent interval or when user navigate to landing page
- SaaS application with high concurrency
- Latency expectation is <200ms
- Realtime and historical data

I want to build…

“ Wish to provide a self service reporting capability to my customers”

aws
COMMUNITY DAY
PUNE 2022

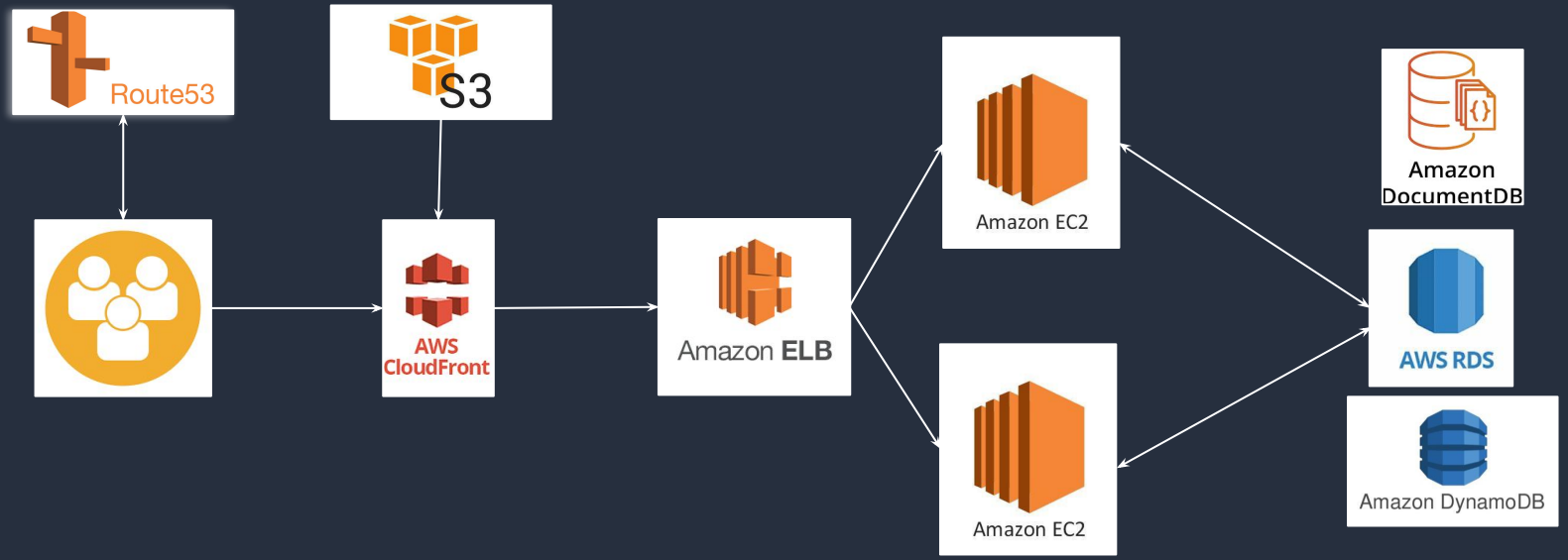
This is what a modern analytics app looks like



An interactive data experience with...

- TBs to PBs of data processed
- 100s to 1000s of queries per sec
- Real-time and historical data analyzed
- Most of the queries are groupby in nature
- Latency requirement < 200ms
- Highly concurrent query workload

Sample Architecture ..



aws
COMMUNITY DAY
PUNE 2022

Sample Architecture ..

@tijothomas214
tijopthomas

Challenges

- Data warehouses including column stores work only with batch ingestion and are not optimal for streaming data.
- DWs are a difficult to scale up and down based on the Query load.
- DWs support low concurrency. Concurrency comes with exponential cost
- DWs has limited support for indexing the data

Demo onboarding data

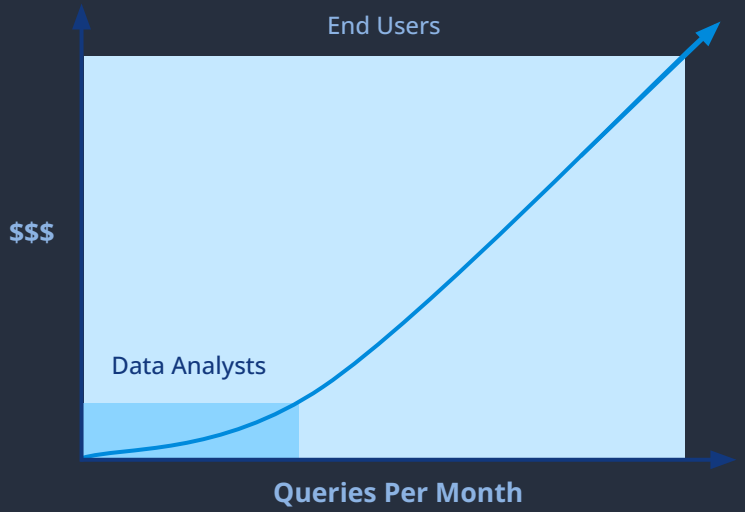
aws
COMMUNITY DAY
PUNE 2022

Demo how to query Druid

aws
COMMUNITY DAY
PUNE 2022

Demo the performance and concurrency

Unlimited Users: Modern analytics apps drive up concurrency



The cost for high concurrency increases exponentially with CDWs

A different database is needed for high concurrency at the best value

Druid shines for modern analytics applications

	General Purpose DBs	Cloud Data Warehouses	Apache Druid
Interactive Analytics at Any Scale	-	-	Yes
High Concurrency (at the Best Value)	-	Yes, but expensive	Yes
Insights on Streaming and Batch Data	-	-	Yes

Recap the requirements

I want to build ...

- High concurrency analytical application
- Wanted to analyze real time and historical data together
- Need latency < 200ms for most of the usecase
- Wanted to scale up or down the infrastructure within reasonable time frame
  eg: <1 hr
- Many other secondary requirements

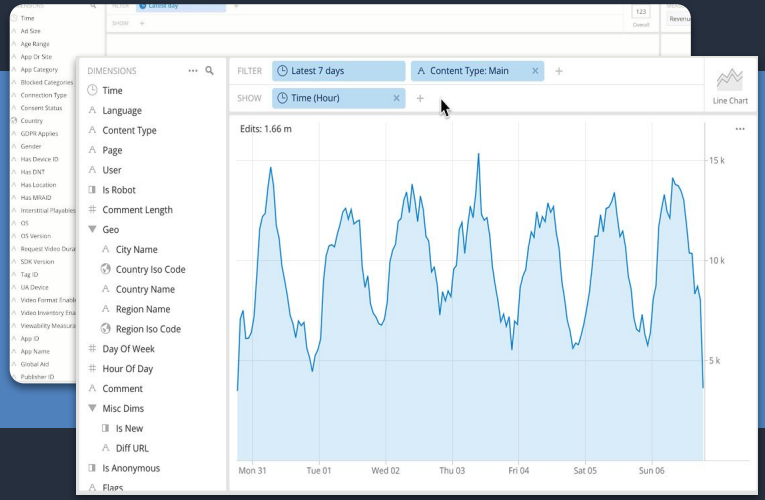
Secondary requirement

Reduce cost per query over older data with higher latency and lesser concurrency

Dont wanted to ingest all raw data but wanted to pre aggregate while ingesting to DB

Let us see how druid helps here

What Is Druid?



Druid is the best database for **modern analytics applications**

10+
Years

400+
Contributors

1,000+
Customers deployed

11,000+
Community members

druid® is...

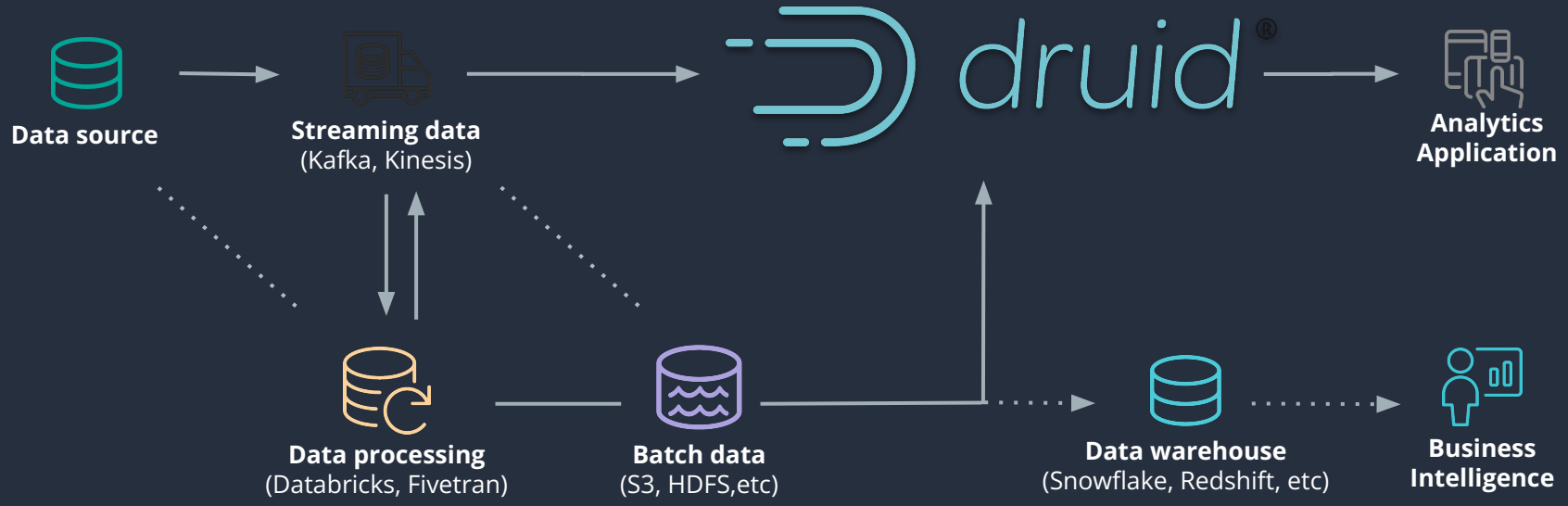
High performance.	*low-latency, distributed query execution and high throughput ingestion*
Real-time.	*event data (clickstream, network flows, user behavior, programmatic advertising, server metrics, IoT…)*
Analytics.	*counting, ranking, statistics...*
Database.	*highly-available, time-sharded, partitioned, columnar, indexed, compressed, versioned materialized view*

Commonly asked questions

Druid a replacement for our existing Data warehouse ?

Where does Druid fit in my big data stack?

Druid in the Data Ecosystem



Commonly asked questions

Is Druid a log aggregation search system

aws
COMMUNITY DAY
PUNE 2022

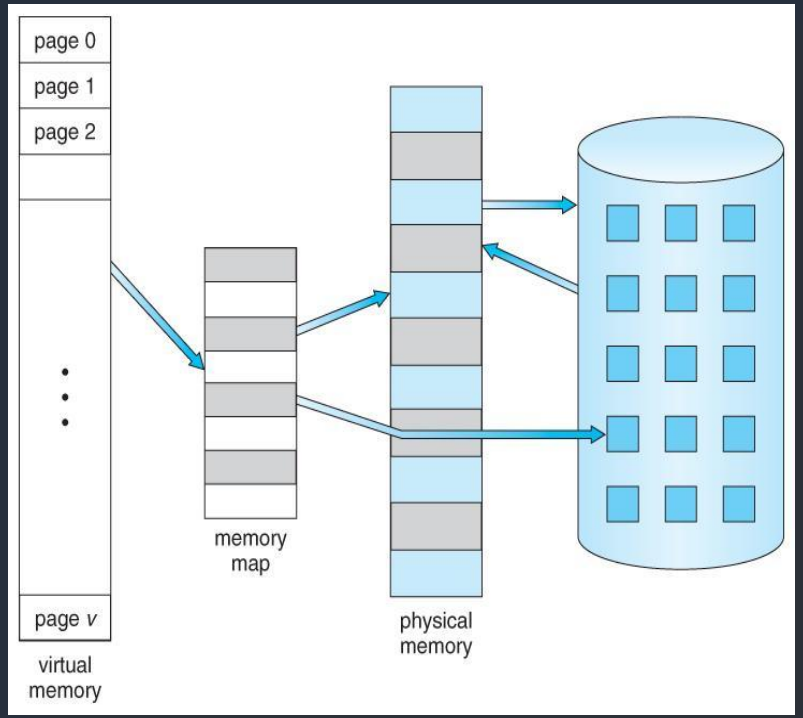
Commonly asked questions

Is Druid a in-memory database?

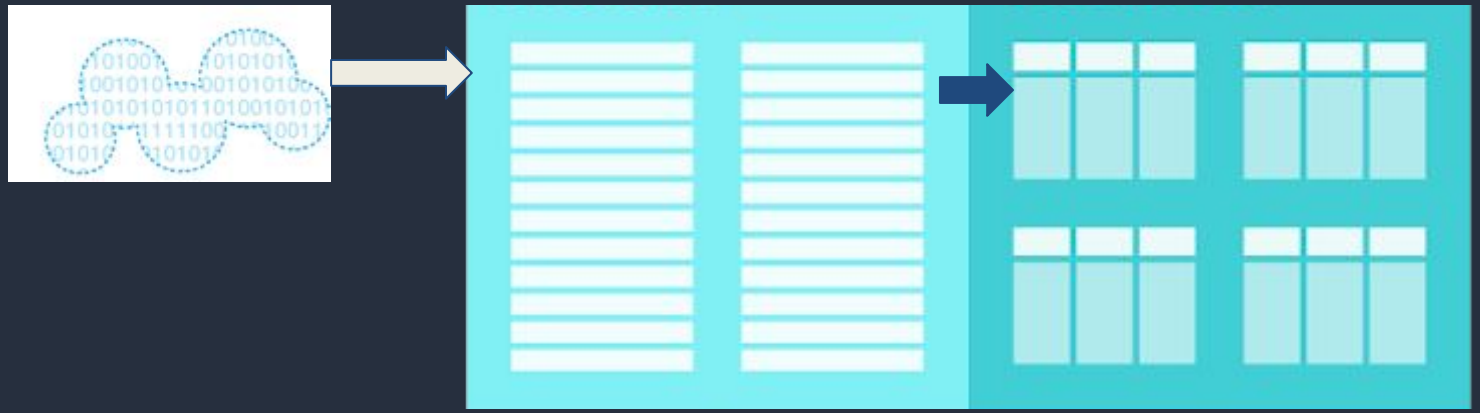
Is Druid a in-memory database?

Query Processing

- Segments can be memory mapped
- Not every time druid need to read from disk



Is Druid a in-memory database?



- While ingesting data will be kept in memory and then written to disk
- The threshold for this is mentioned in the ingestion specification

imply &

druid®

Imply Community Team
community@imply.io

Imply Training Program
https://learn.imply.io

Druid User Forum by Imply
https://www.druidforum.org

Druid Professionals Group
https://www.linkedin.com/groups/8791983/

Druid Meetups
https://www.meetup.com/pro/apache-druid/

Druid News & Info
@druidio #apachedruid @implydata

Druid Slack
#druid
https://druid.apache.org/community/join-slack

Druid @ Imply
https://imply.io/what-is-druid/

aws
COMMUNITY DAY
PUNE 2022

Q&A

aws

COMMUNITY DAY

PUNE 2022

Thank you!

See you at the AWS Community Day Pune 2023