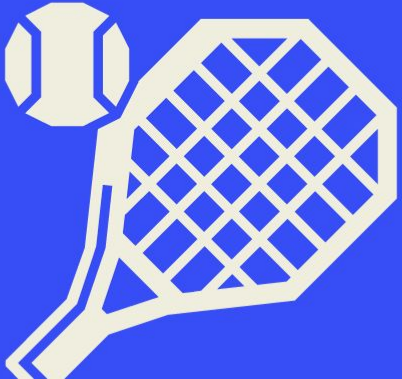
Harnessing Real-Time Power: Inside Sofascore’s Data-Driven Infrastructure

October, 2024: Nog.hr



SOFASCORE — FOR ALL SPORTS. SPORTS FOR ALL.

Sofascore

TABLE OF CONTENTS

INTRODUCTION

Karlo Knežević

Husband

Father

35 years old, from Zagreb

Sofascore, Faculty of Electrical Engineering and Computing, Algebra University

PhD. in Machine Learning and Evolutionary Computation applying to symmetric cryptography

Head of AI @ Sofascore

01

What is Sofascore

25

SPORTS

10
10
20
8
4

Reebok
FORTRESS AXALUX
Reebok
SJ
SVENSKA SPEL
Beckers
nya · hem
Beckers
nya · hem
32
SPORTEWA
26
eebok
SPORTEWA

UEFA
COUPE DES CLUBS CHAMPIONS
EUROPÉENS

NK HRVATSKI BOJOVNIK
MOKRICE
2
RADNIK
9
9

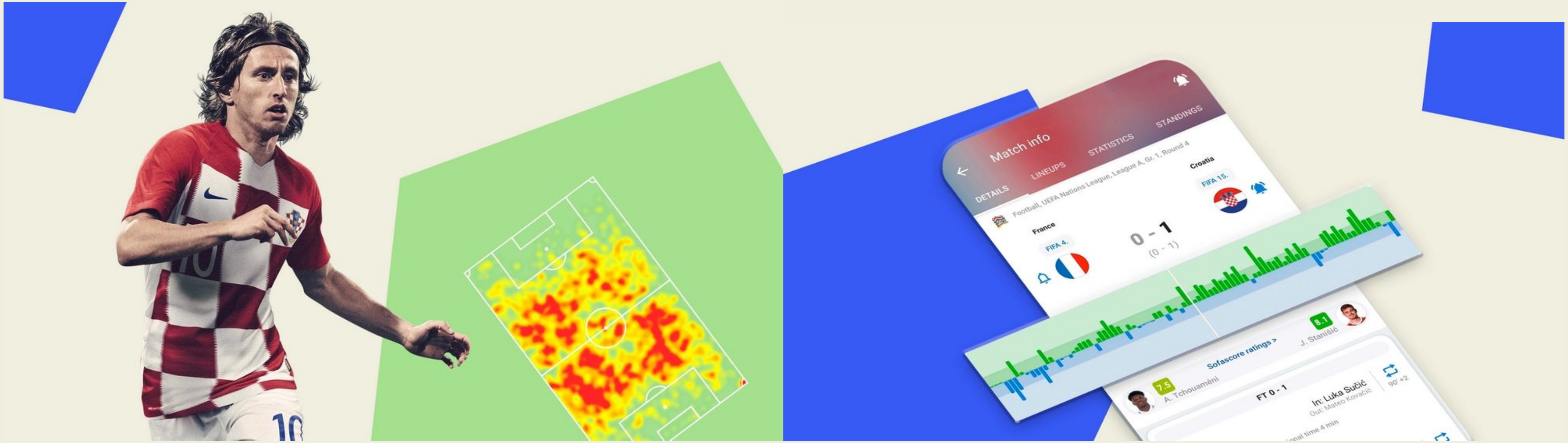
28

MILLION USERS

02

Data in Sofascore

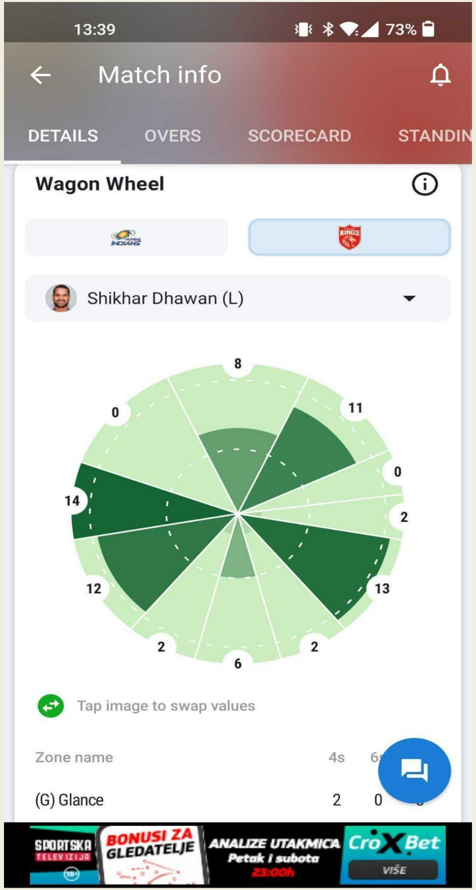
SPORTS DATA

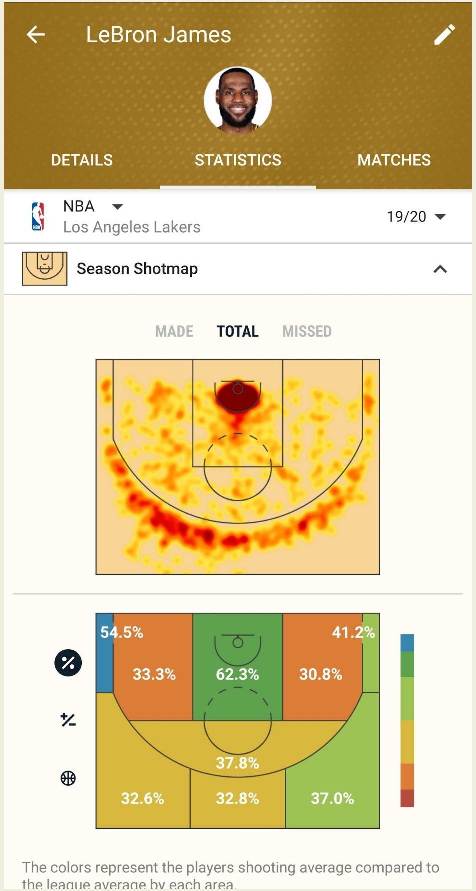

Match info
DETAILS
LINEUPS
STATISTICS
STANDINGS
Football, UEFA Nations League, League A, Gr. 1, Round 4
France
FIFA 4.
0 - 1
(0 - 1)
Croatia
FIFA 15.
Sofascore ratings >
7.5
A. Tchouaméni
8.1
J. Stanišić
FT 0 - 1
In: Luka Sučić
Out: Mateo Kovačić
90' +2

DATA IN SOFASCORE

Sports data in Sofascore

- 25 sports (for now)
- Data on players, teams, leagues, events (matches), coaches, referees...
- Insightful statistics
- Player rating
- Graphs and visualizations





The colors represent the players shooting average compared to the league average by each area

Where does this data come from?

- Buying providers' data (Opta Sports)
- Scraping data (web)
- Manual inserting (Sofascore Data Team)
- Crowdsourcing

100GB

How much data are we talking about?

TABLE	ROWS
event	~6,3M
player	~900k
team	~350k
tournament	~17k
...	...

USER ACTIONS DATA

User actions data in Sofascore

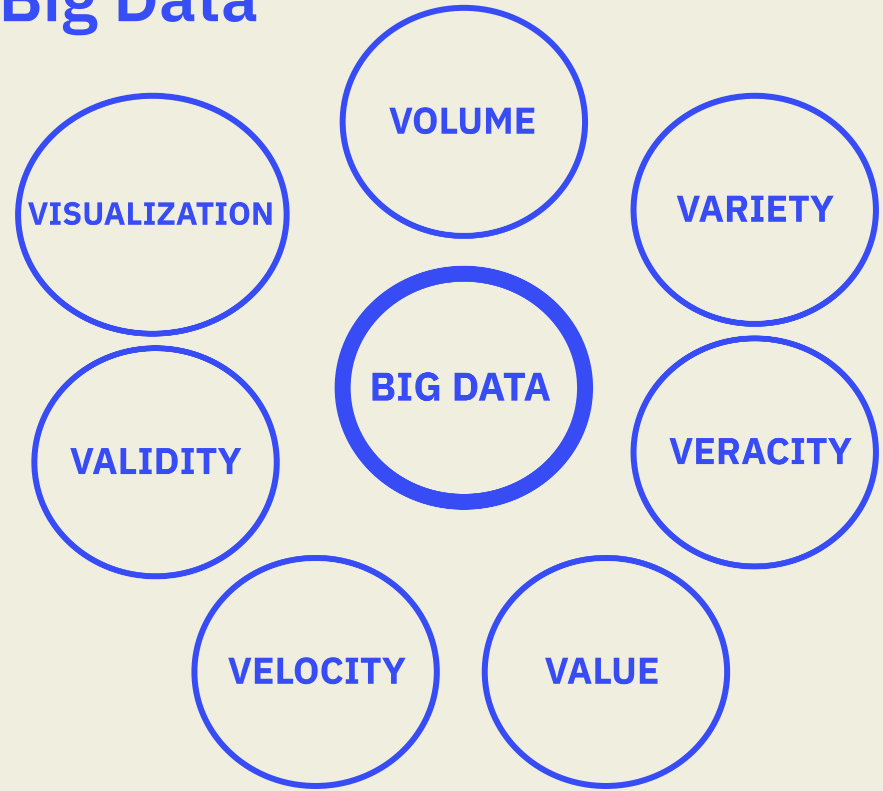
- ❗ **Clickstream** - records of user's actions through their app journey
- +200 actions (events)
- **Event** - important occurrence in our app that we measure
- e.g. follow_player, open_league, open_player, ad_click, app_remove, ...

Big Data

“extremely large dataset that may be analysed computationally to reveal patterns, trends, and associations, especially relating to human behaviour and interactions”

1PB

❗ The 7 Vs of Big Data



BIG DATA

VOLUME

VARIETY

VERACITY

VALUE

VELOCITY

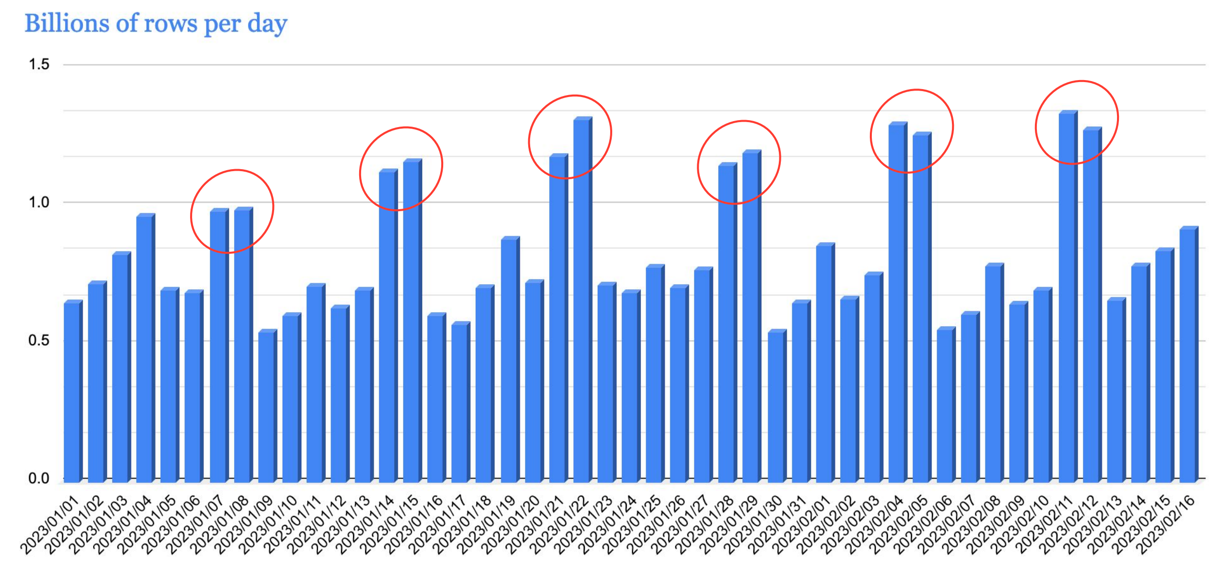
VALIDITY

VISUALIZATION

How much data are we talking about?

VOLUME

- ~1.5PB of user actions data since February 2019
- 1 trillion rows, 268 columns in **bq.events** table in **ClickHouse**
- ~700GB arriving daily



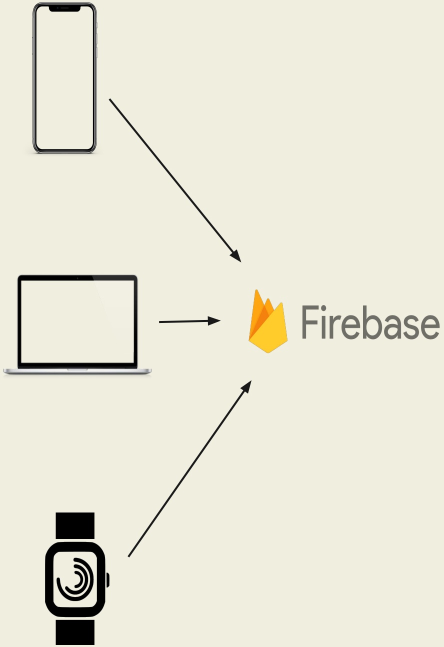
❗ Data Warehouse (DWH)

- A place where all enterprise data from multiple sources is consolidated into a single source of truth
- Finance, marketing, product and users data integrated into one database
- Used by data scientists and BI in order to make data-driven decisions
- Sofascore’s DWH - ClickHouse

03

Data Pipeline

DATA PIPELINE



Firebase

"Firebase is an app development platform that helps you build and grow apps and games users love. Backed by Google and trusted by millions of businesses around the world."

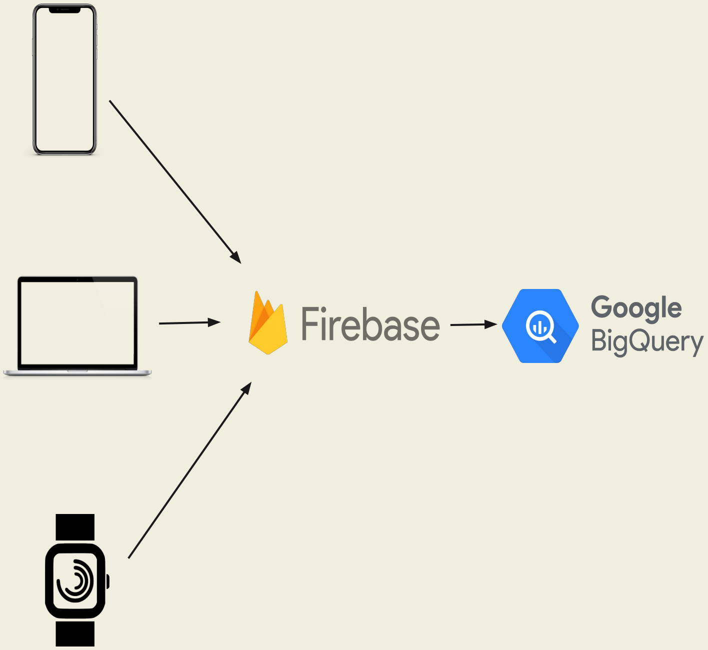
- Offering services like:
    - **Analytics**, Authentication, Databases, File Storage, Push Messages, ...

Google Analytics for Firebase

- Reporting up to 500 different types of events
- Associate up to 25 parameters with each event
- Data is exported daily to **BigQuery**
    - Duplicates occurring (due to client's network issues)
    - < 2% duplicates

VERACITY

DATA PIPELINE

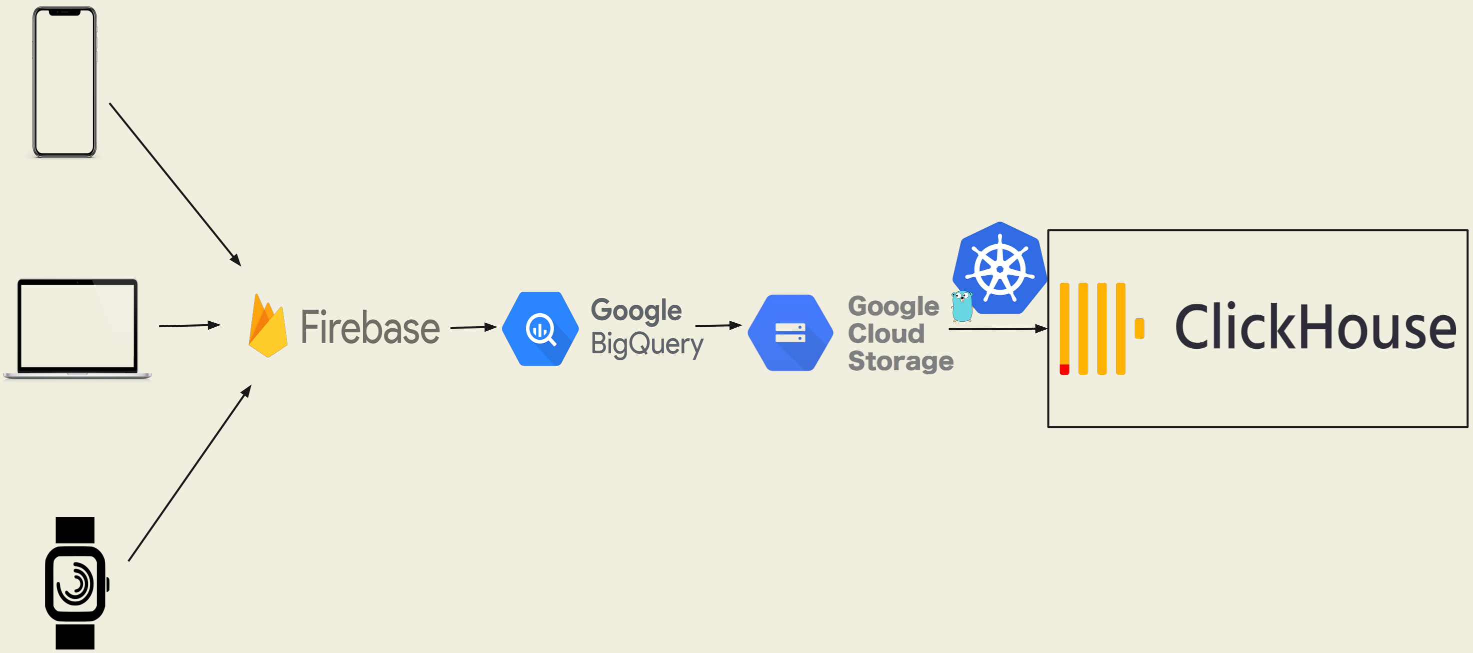


Google BigQuery

BigQuery is a completely serverless and cost-effective enterprise data warehouse. It has built-in machine learning and BI that works across clouds, and scales with your data.

- Table named *events_YYYYMMDD* is created each day within BQ dataset
- Each column in that table represents an event-specific parameter
- Apart from event-specific parameters other columns are:
    - Event (event_name, event_date, event_timestamp...)
    - User (user_pseudo_id, user_id, user_first_touch_timestamp)
    - Geographical (continent, country, region...)
    - Device (category, language, operating_system, ...)
    - App info (id, version, install_source, ...)
    - Traffic Source (name, medium, ...)

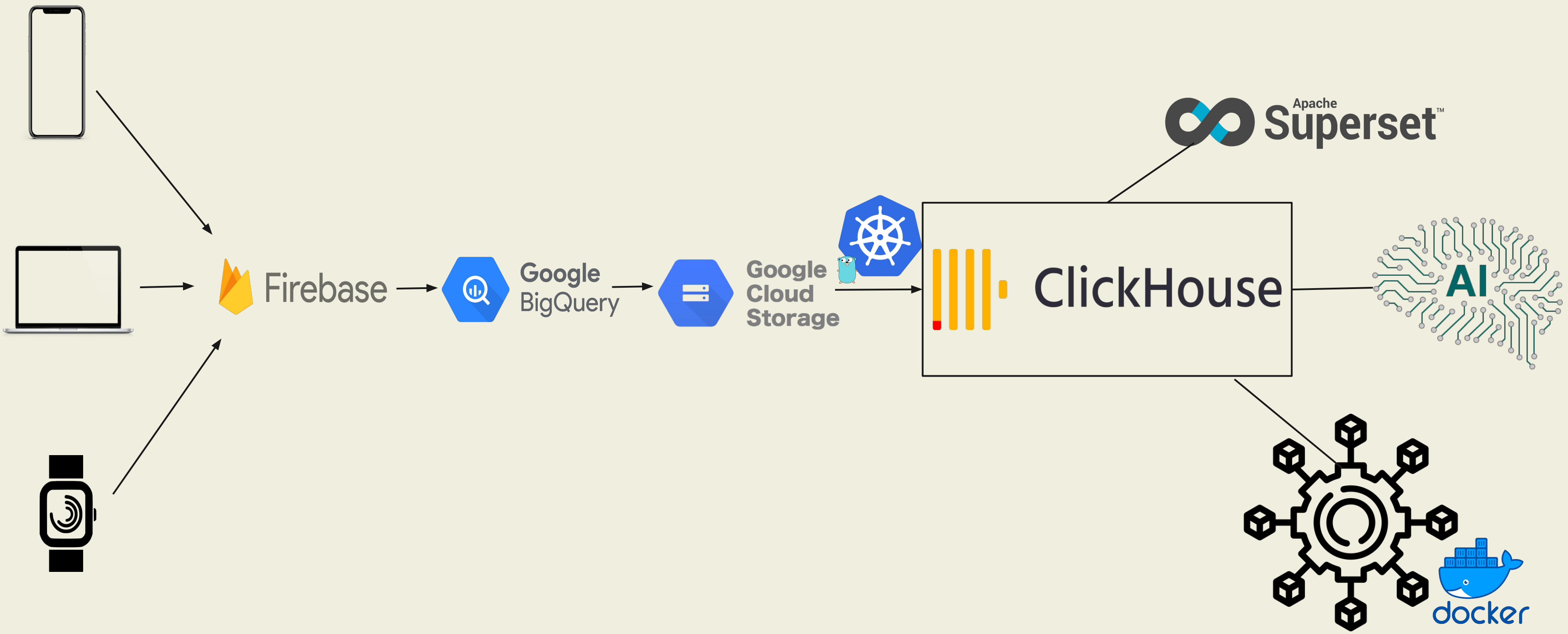
DATA PIPELINE



- **Kubernetes** (**K8s**) - open-source system for automating deployment, scaling, and management of containerized applications
- Dockerized Go script that imports data into Clickhouse from files exported to GCS from BigQuery
- Parallel import on 15 servers in Sofascore Kubernetes cluster
- Import using K8s cluster speeds up the import ~15 times 🚀

VELOCITY

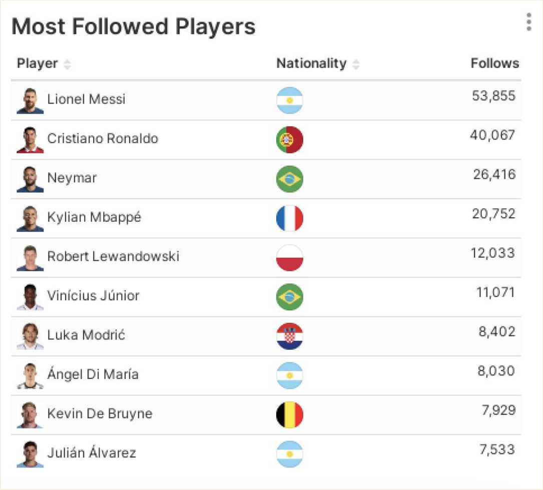
DATA PIPELINE



VISUALIZATION

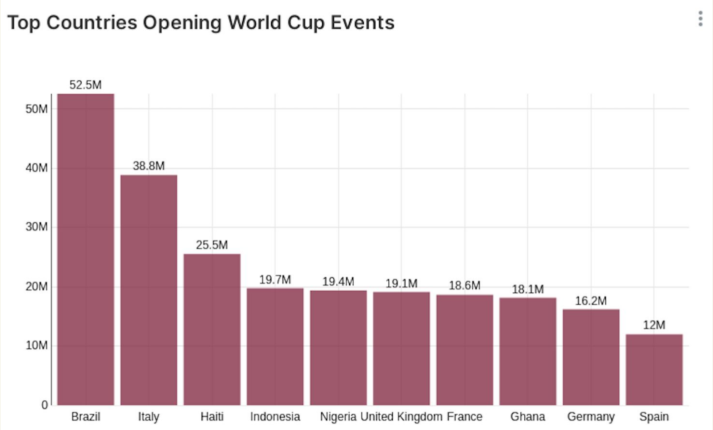
Apache Superset™

- Most popular (by Github ⭐) **open-source** BI and analytics visualization platform in the world
- Software application for data exploration that handles data at petabyte scale
- Wide range of database support



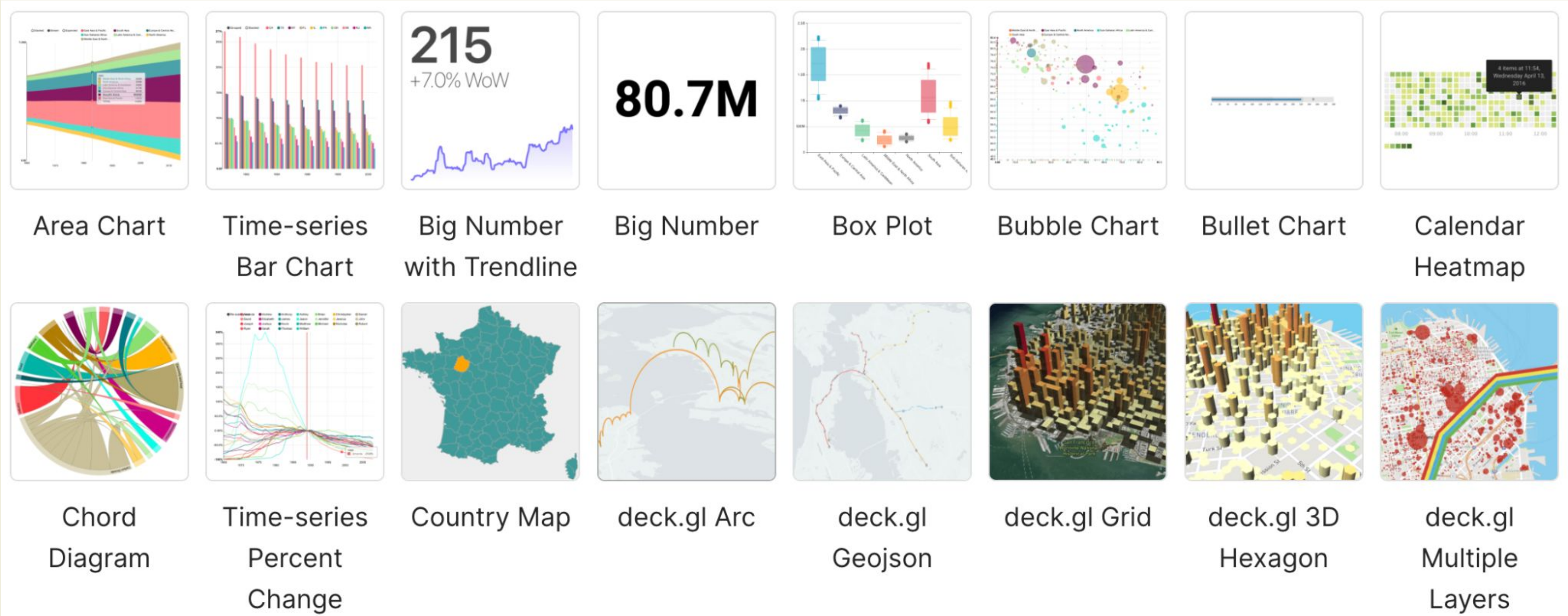
Most Followed Players

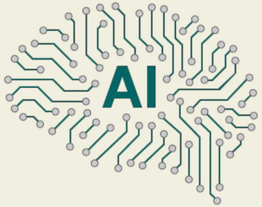
Player	Nationality	Follows
Lionel Messi		53,855
Cristiano Ronaldo		40,067
Neymar		26,416
Kylian Mbappé		20,752
Robert Lewandowski		12,033
Vinícius Júnior		11,071
Luka Modrić		8,402
Ángel Di María		8,030
Kevin De Bruyne		7,929
Julián Álvarez		7,533



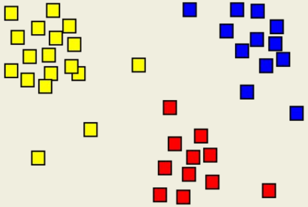
Top Countries Opening World Cup Events

Apache Superset™



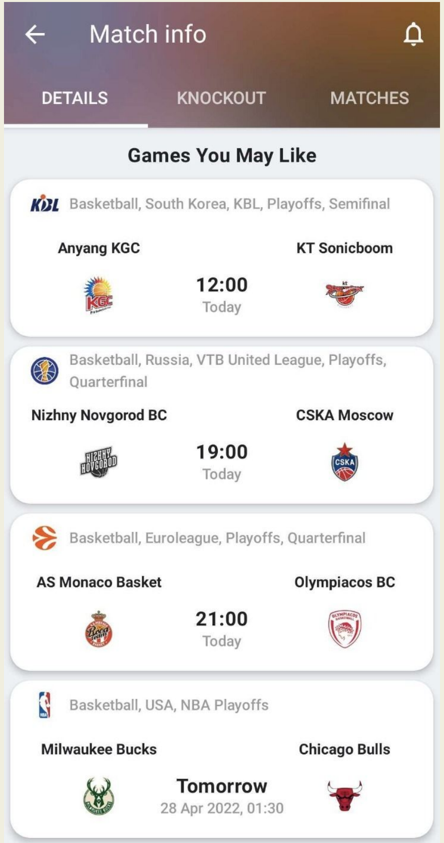


- User clustering



- Predictions (anticipate change in market)

- Content recommendation (user retention)



Games You May Like

Basketball, South Korea, KBL, Playoffs, Semifinal
Anyang KGC 12:00 Today KT Sonicboom

Basketball, Russia, VTB United League, Playoffs, Quarterfinal
Nizhny Novgorod BC 19:00 Today CSKA Moscow

Basketball, Euroleague, Playoffs, Quarterfinal
AS Monaco Basket 21:00 Today Olympiacos BC

Basketball, USA, NBA Playoffs
Milwaukee Bucks Tomorrow 28 Apr 2022, 01:30 Chicago Bulls

VALUE

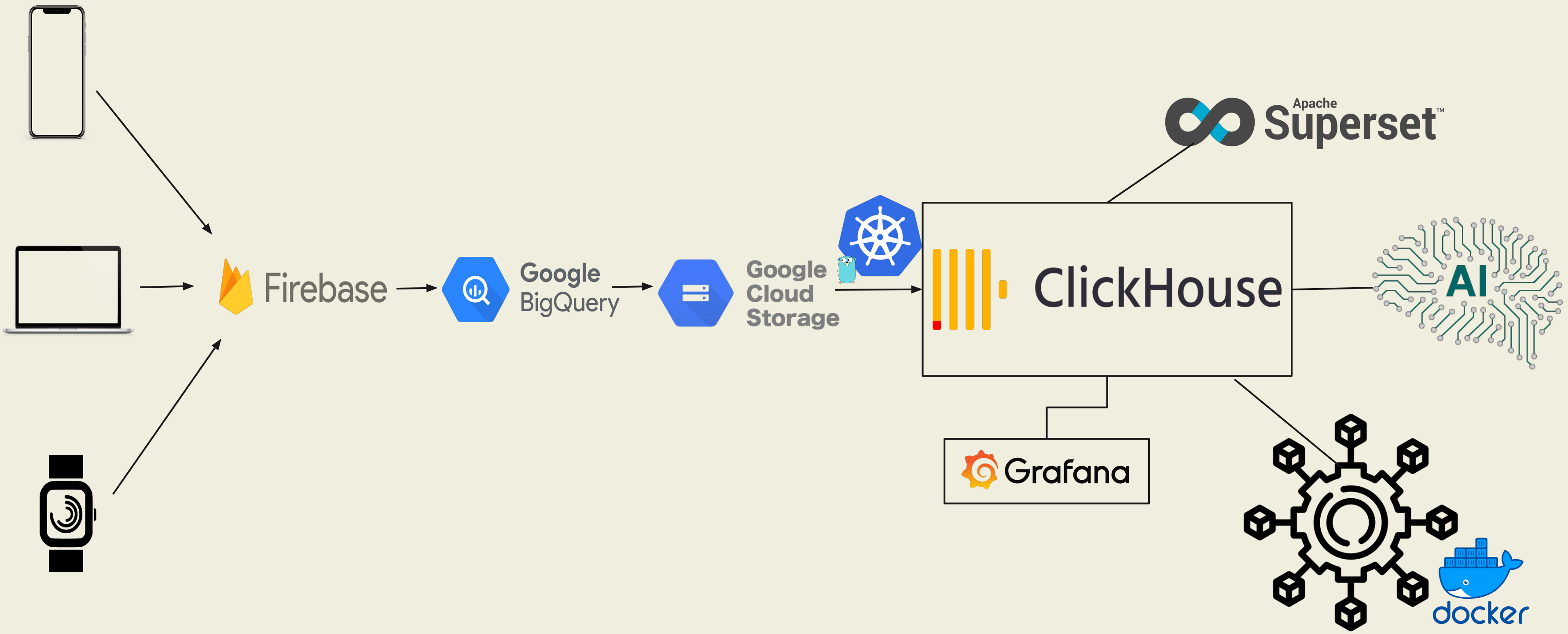
Sofascore

docker

- Dockerized microservices written in Go and Python
- Defined as different resources inside the kubernetes cluster (cronjobs, jobs, deployments…)
- Importing sales and financial data from Google Sheets and Smartsheets
- Importing data from APIs

VARIETY

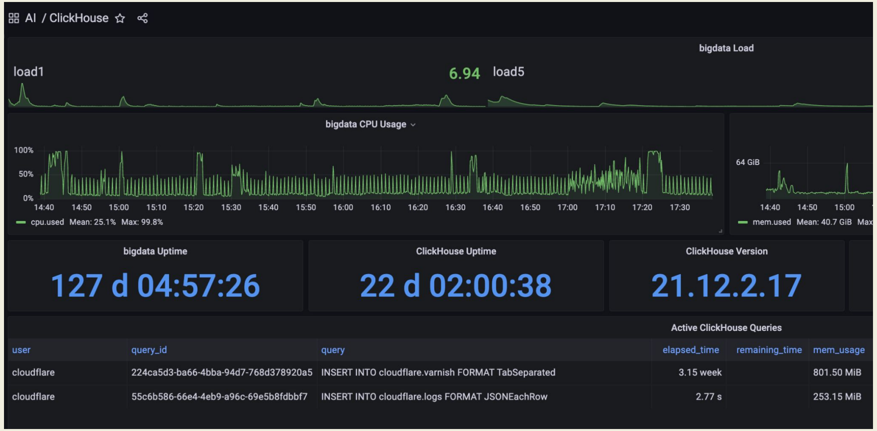
DATA PIPELINE



Grafana

- Open source application for multi-platform analytics and interactive visualization
- Allows you to query, visualize, alert on and understand your metrics
- Well suited for time series data visualization
- **Monitoring**:
  - Services
  - Clickhouse performance
  - Server load

VALIDITY



bigdata Uptime
127 d 04:57:26

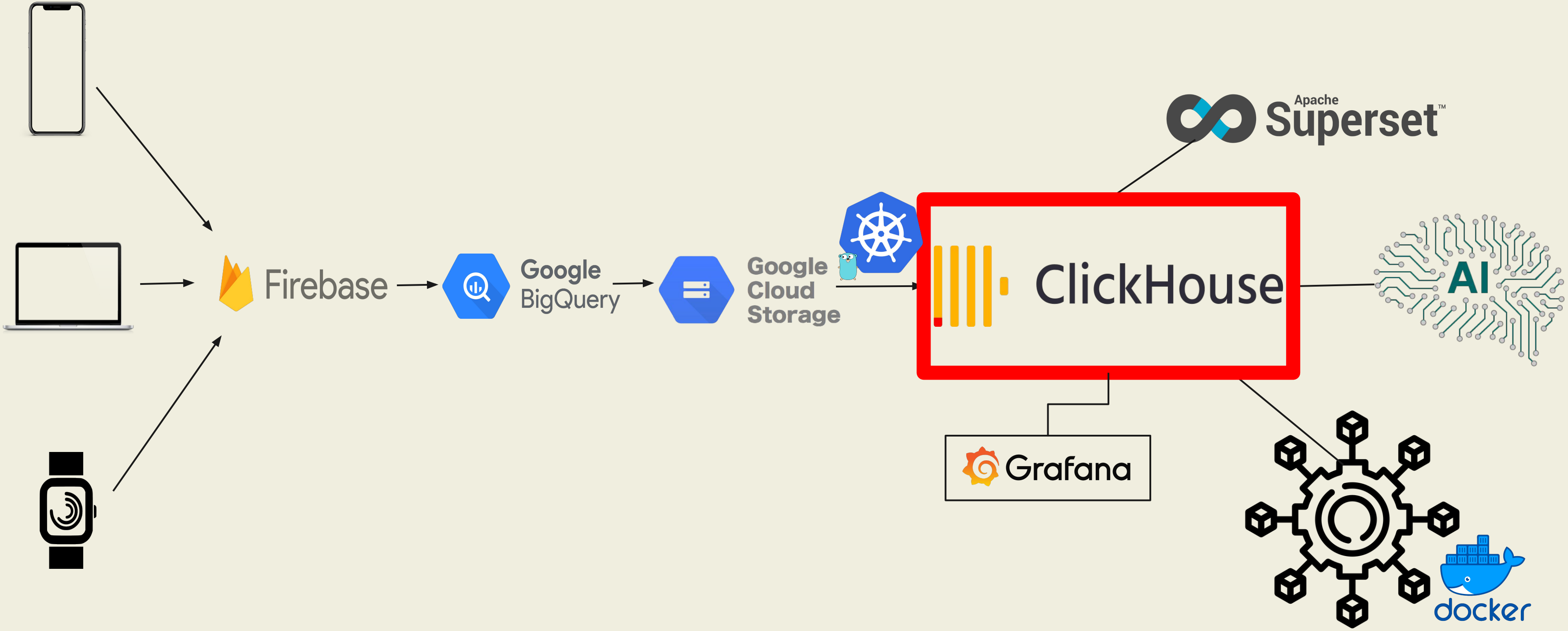
ClickHouse Uptime
22 d 02:00:38

ClickHouse Version
21.12.2.17

Active ClickHouse Queries

user	query_id	query	elapsed_time	remaining_time	mem_usage
cloudflare	224ca5d3-ba66-4bba-94d7-768d378920a5	INSERT INTO cloudflare.varnish FORMAT TabSeparated	3.15 week		801.50 MiB
cloudflare	55c6b586-66e4-4eb9-a96c-69e5b8fdbbf7	INSERT INTO cloudflare.logs FORMAT JSONEachRow	2.77 s		253.15 MiB

DATA PIPELINE



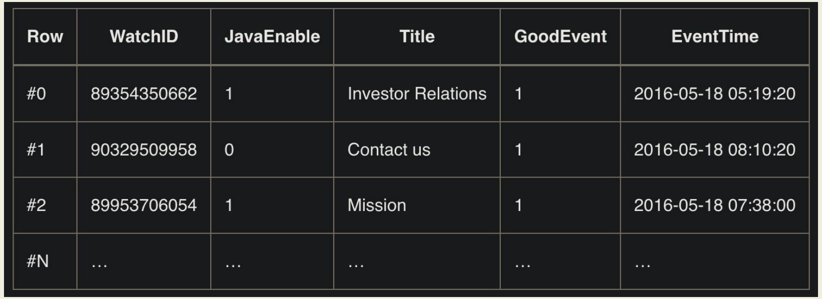
ClickHouse

- **Click**stream + Data Ware**house** = **ClickHouse (CH)**
- Open source **column-oriented** SQL database management system (DBMS) for online analytical processing (**OLAP**)
- Fist developed at Yandex and launched in production in 2012 to power **Yandex.Metrica**

Yandex Metrica

Row oriented DBs

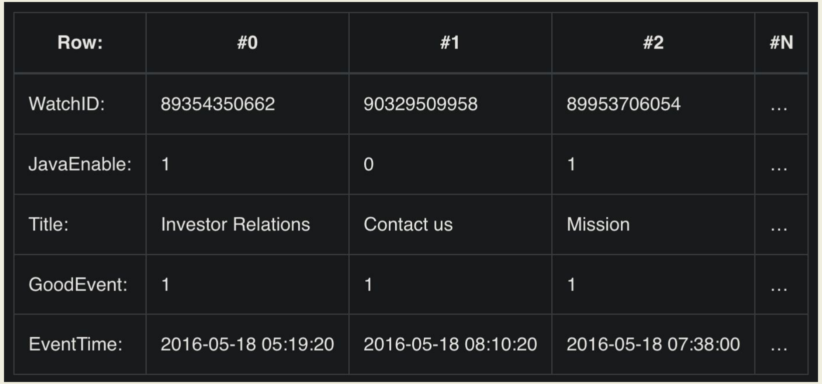
- Data associated with a **record** next to each other in memory
- Common row oriented DBs (Postgres, MySQL, MariaDB...)



Row	WatchID	JavaEnable	Title	GoodEvent	EventTime
#0	89354350662	1	Investor Relations	1	2016-05-18 05:19:20
#1	90329509958	0	Contact us	1	2016-05-18 08:10:20
#2	89953706054	1	Mission	1	2016-05-18 07:38:00
#N	…	…	…	…	…

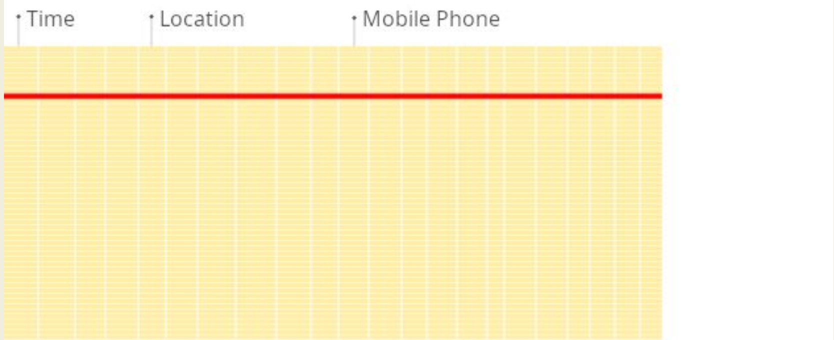
Column oriented DBs

- Data associated with a **field** next to each other in memory
- Common column oriented DBs (Redshift, BigQuery, ClickHouse...)

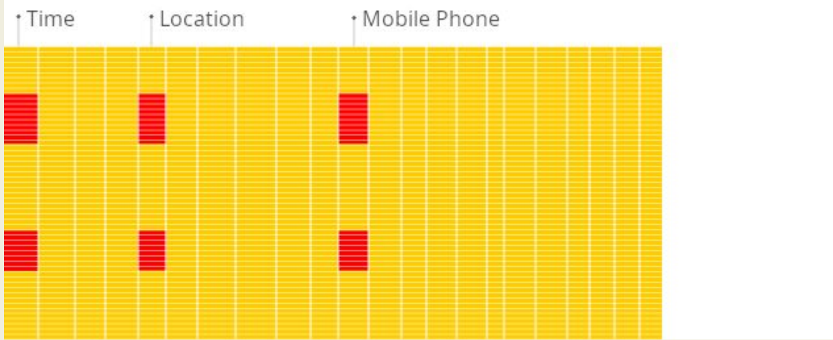


Row:	#0	#1	#2	#N
WatchID:	89354350662	90329509958	89953706054	…
JavaEnable:	1	0	1	…
Title:	Investor Relations	Contact us	Mission	…
GoodEvent:	1	1	1	…
EventTime:	2016-05-18 05:19:20	2016-05-18 08:10:20	2016-05-18 07:38:00	…

Row oriented DBs



Column oriented DBs

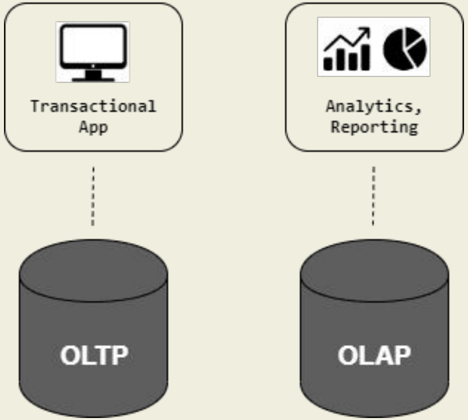


OLTP

- Online **Transaction** Processing
- Most row oriented DBs
- Many users performing varied queries and updates
- SQL primary language for interaction

OLAP

- Online **Analytical** Processing
- Most column oriented DBs
- Fewer users performing deep data analysis
- Utilizes particular query language other than SQL



Clickhouse features

Parallel query processing

Multiple table engines

OPEN SOURCE

Cost effective

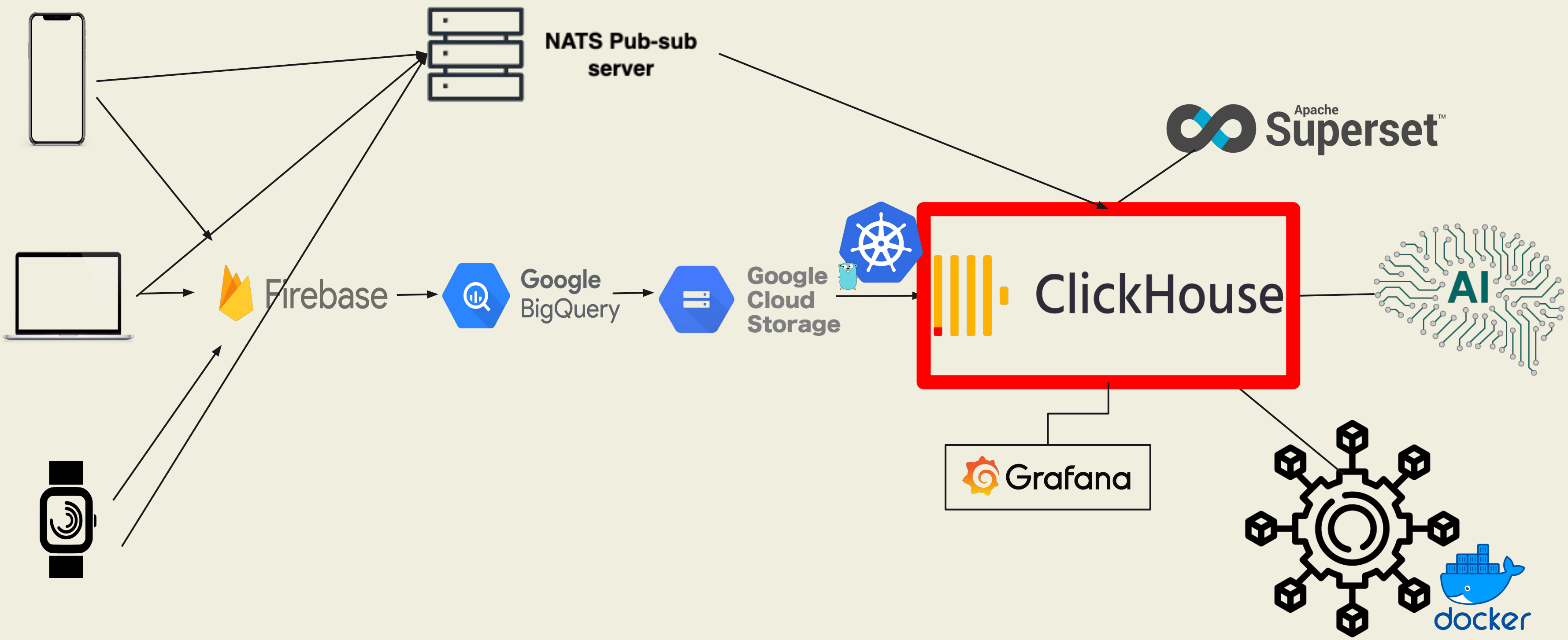
$f(x)$

HUGE number of functions

Compression

Index support

NATS & REAL-TIME ANALYTICS AND AI

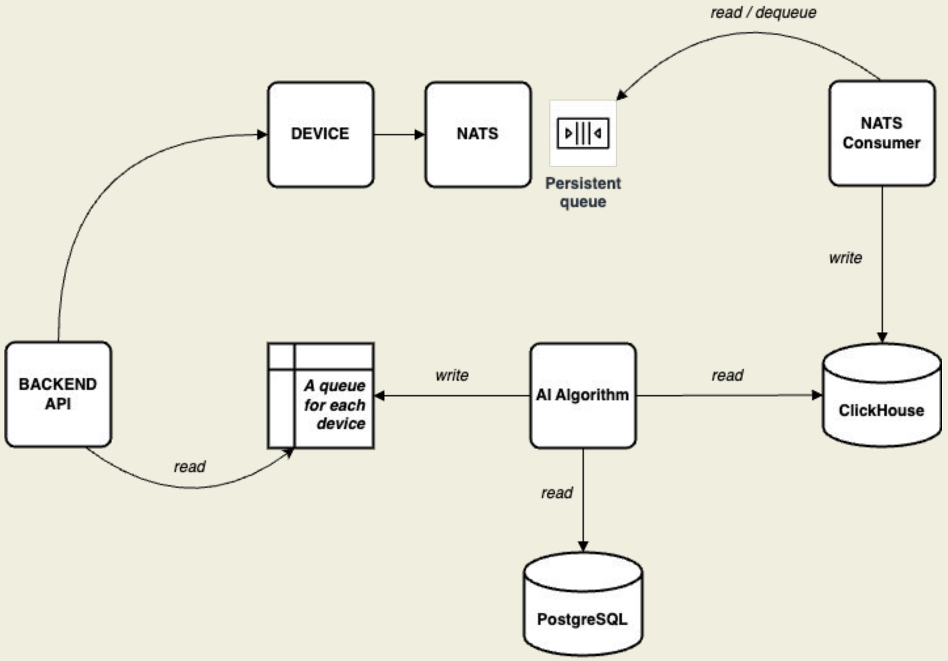


NATS Pub-sub server

Sofascore

NATS & REAL-TIME ANALYTICS AND AI

NATS

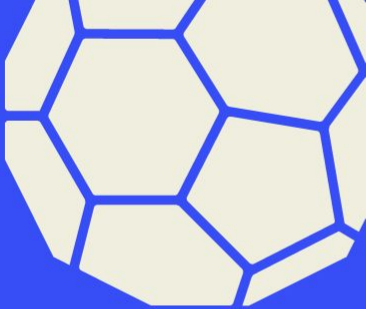


04

Summary

Summary

- ➔ No. 1 sports platform in the world
- ➔ Big Data: challengeable
- ➔ User action data > sports data
- ➔ Data latency
- ➔ Real-time data pipeline with NATS



Thank you!

karlo.knezevic@sofascore.com