



OPENWORKS

SKYSQL SERVERLESS ANALYTICS POWERED BY SPARKSQL - PART 1

WILL START SHORTLY.



OPENWORKS

SKYSQL SERVERLESS ANALYTICS POWERED BY SPARKSQL - PART 2

WILL START SHORTLY.



OPENWORKS

BE UNSTOPPABLE



OPENWORKS

SERVERLESS ANALYTICS

SARAVANA KRISHNAMURTHY, VICE PRESIDENT, MARIADB

OBJECTIVE

1. Why Serverless Analytics?
2. What problem are we solving?
3. There are so many cloud analytics services available (BigQuery, Snowflake, Athena, Azure Synapse, Dremio, RockSet.. How can we compare?
4. What's unique about Serverless Analytics powered by Apache Spark?

THE DB INDUSTRY CONTINUES TO CHANGE...



- Moore's Law
- Cloud
- Cheap Storage



- OLAP Cubes
- ETL
- Materialized Views
- Kimball, Inman, Datavault
- Schema on write
- Data Warehouse
- Server provisioning
- Data Engineers

WHERE ANALYTICS IS HEADING?

- OLAP Cubes
- ETL
- Materialized Views
- Kimball, Inman, Datavault
- Schema on write
- Data Warehouse



Lakehouse/DataMesh

- Server provisioning



Serverless

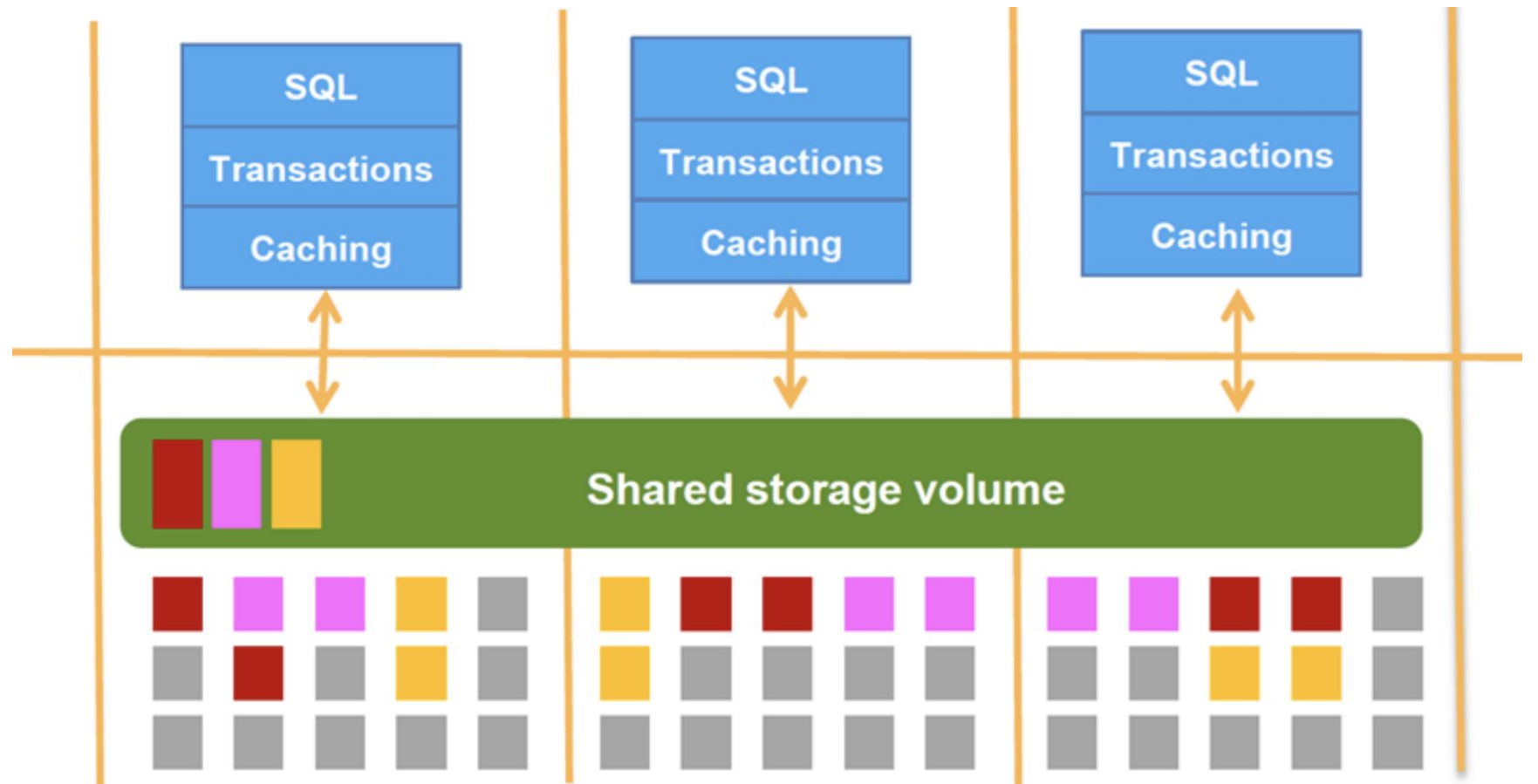
Role of CDO in large companies*: 12% in 2012 -> 65% in 2021

*Source: <https://venturebeat.com/business/despite-high-demand-for-data-leadership-cdo-roles-need-improvement/>

MODERN DATABASES ARE TRANSLYTICAL

Aurora Parallel Query

- Limited concurrent parallel sessions
 - Max of 16 (db.r4.16xlarge)
- OLTP interference

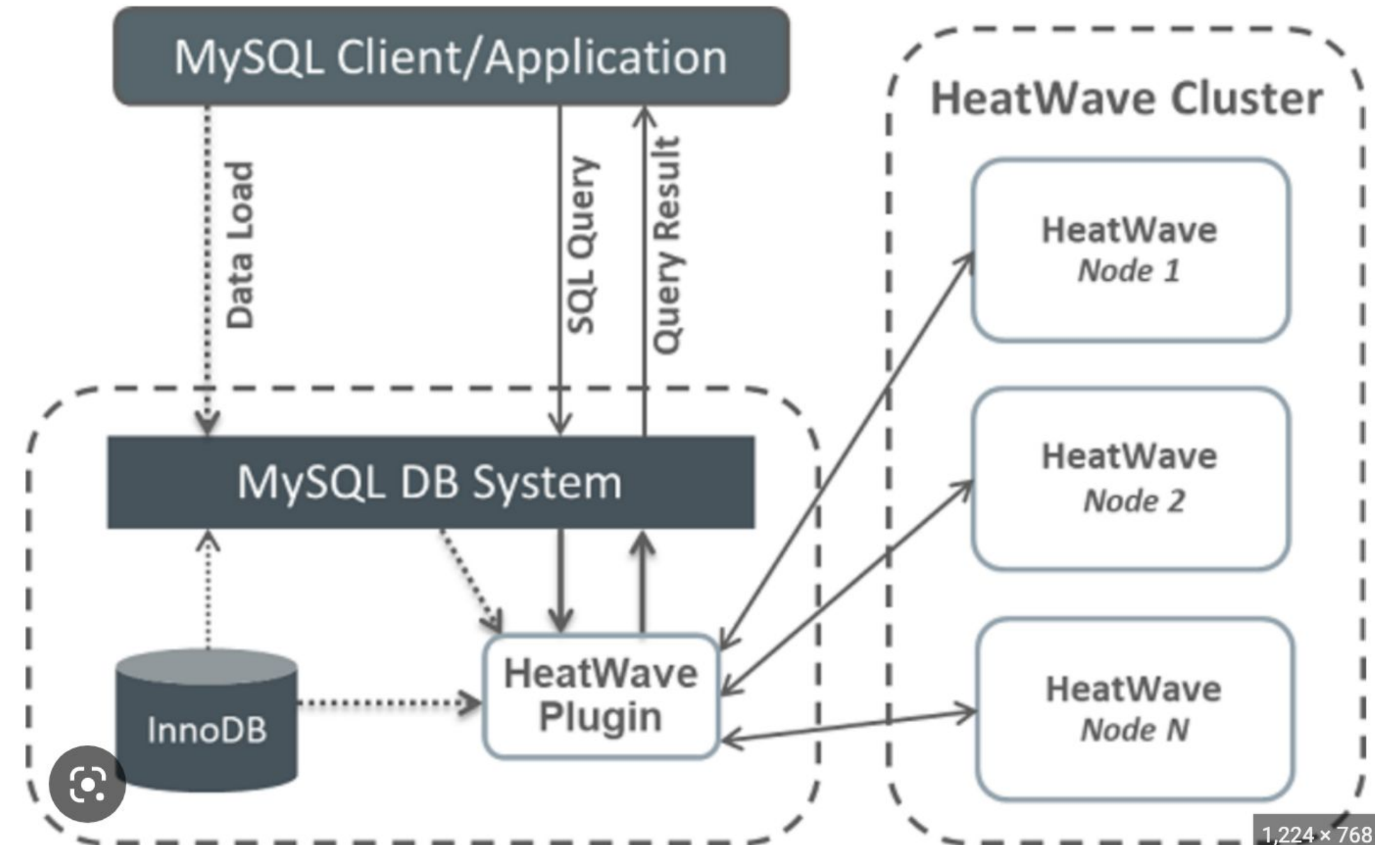


MODERN DBS ARE TRANSLYTICAL

Oracle Heatwave

- Disjointed analytics engine
 - In-memory, pre-provisioned
 - **Expensive**
- MySQL syntax

SAP Hana, SingleStore are all in-memory, **very expensive on cloud**



Concurrency remains a huge challenge for all



WHY SERVERLESS ANALYTICS IN SKYSQL?

- Achieve Interactive analytics on large operational data sets
 - Any Query - Low latencies for large scale aggregations, complex queries or point operations
 - And, without impeding concurrently running OLTP apps (Analytic interference)
 - *How: Delegate to row-oriented Btree indexes for point and range lookups. And, use vectorized code execution on in-memory columnar data for large scale aggregations or complex queries. Use isolated compute for Analytics*
- Achieve predictable response time for concurrent users/Apps
 - e.g. BI users running ad hoc queries on large data sets or Embedded in-app Analytics
 - *How: Compute ballooning - scheduler dynamically acquires more CPU from cloud provider*
- Blend with external sources and any format
 - Join data in other repositories like S3, Warehouses and SQL DBs.
 - Work with data in any format without ETL : Json, CSV, Parquet, SQL, etc

WHY SERVERLESS ANALYTICS

Zero-ETL - operate on Fresh data without data pipelines to a warehouse

Little or no OLTP interference

Serverless - no provisioning, borrow capacity just when required

Pay for use - Pay only for the compute consumed

Support high concurrency without pre-provisioning (cheap)

Enables **advanced analytics** on data stored in Data lakes and SkySQL Cloud databases



SERVERLESS ANALYTICS - A MANAGED SKYSQL SERVICE

Frictionless, Real time Analytic Query Service



Apps



Notebooks

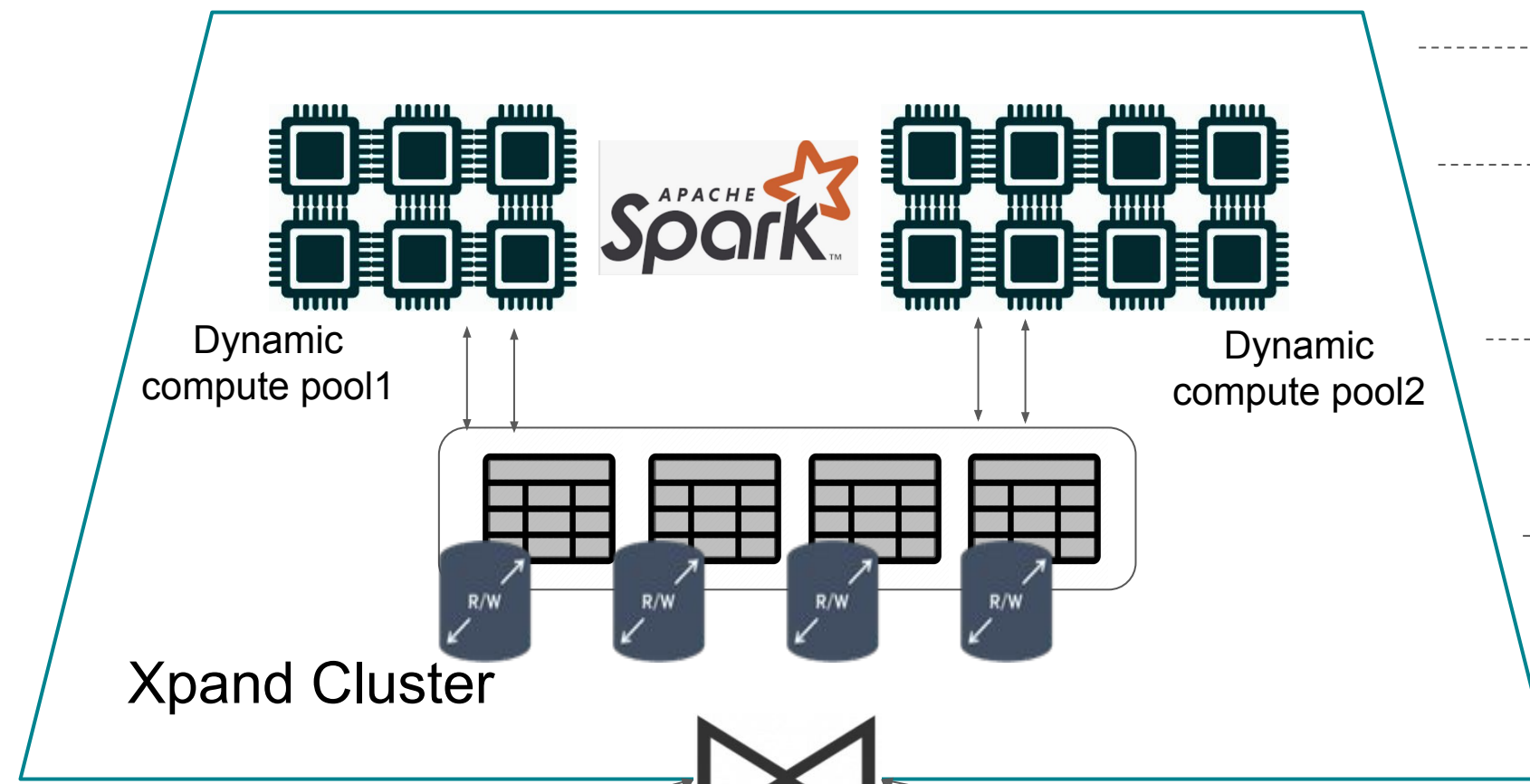
(SQL, Java, Python, Scala, R)

On-demand compute
*(Pay only for what you use
Elastically scale to 1000's of nodes)*

Apache Spark optimized for Xpand

Cache data in Spark with built-in real time CDC

Parallel analytics on any data
Join, aggregate to other data sources



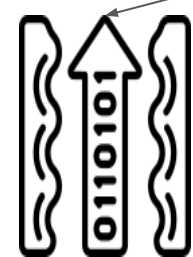
Schema inference

Nested data support

Fast: vectorized, HashJoin, Aggregation

Parallely connect Spark compute partitions to Xpand

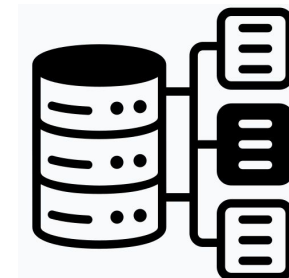
Push down query fragments to Xpand Columnar, Layered index



Streams



Data lakes



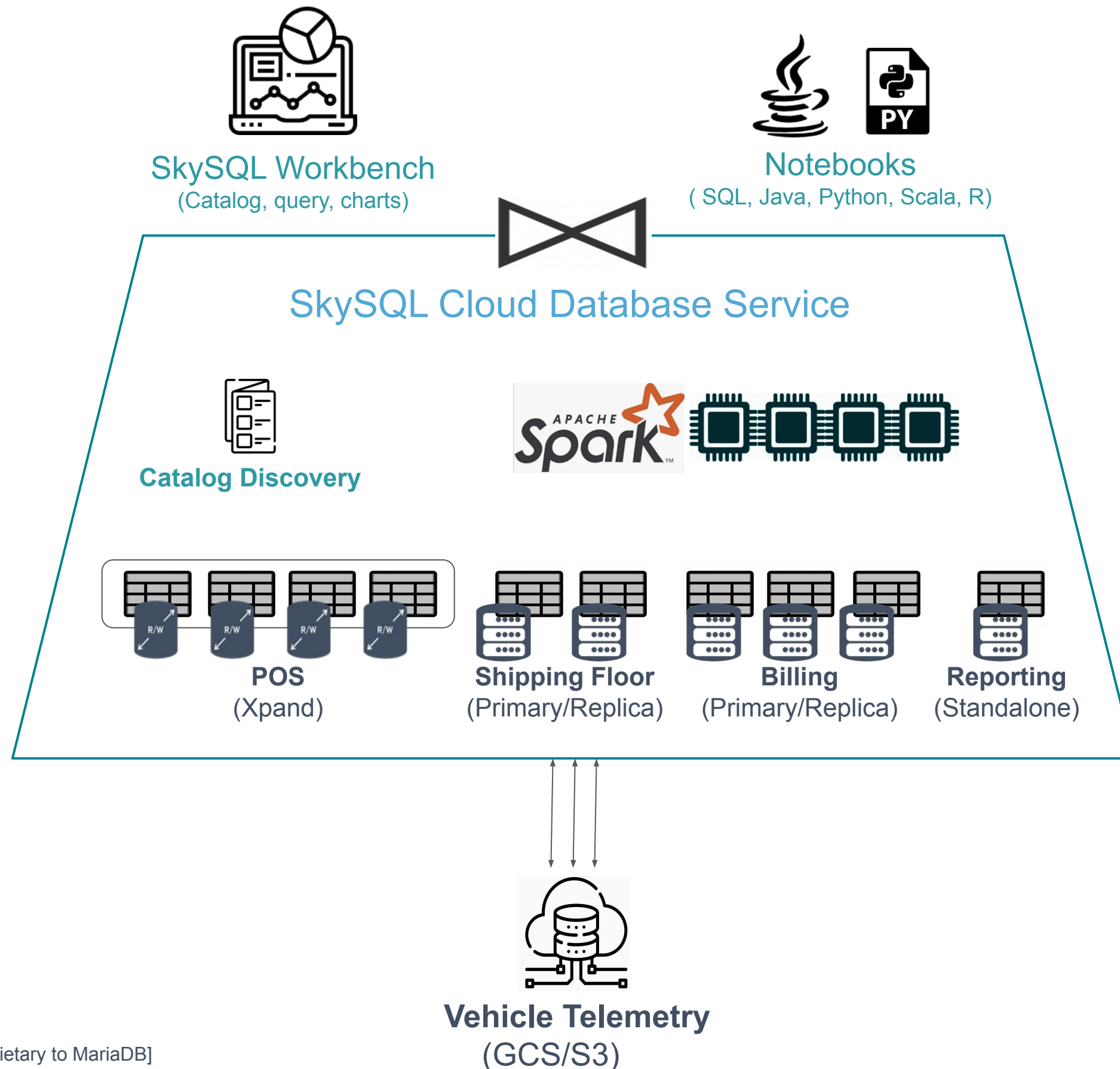
Warehouses, Any RDBMS



S3, GCS ...
(parquet, Json, CSV ...)

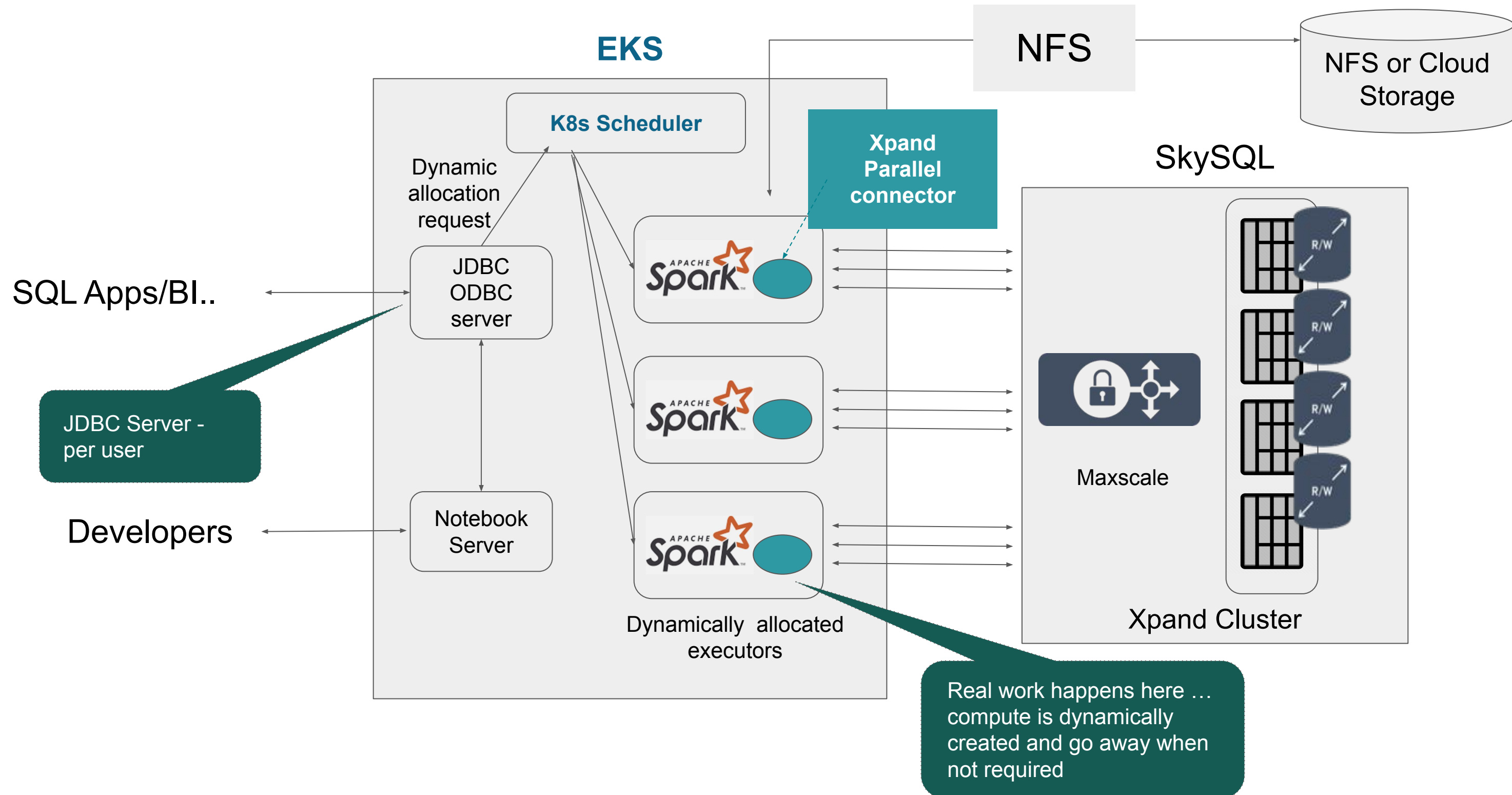


COMMON USE CASE - REPORTING ON ONLINE DATABASES



- Whitebox - an e-commerce platform fulfills online orders through Amazon
- Several online databases capture transactions, shipping, Billing...
- Offer insights to customers through Analytics

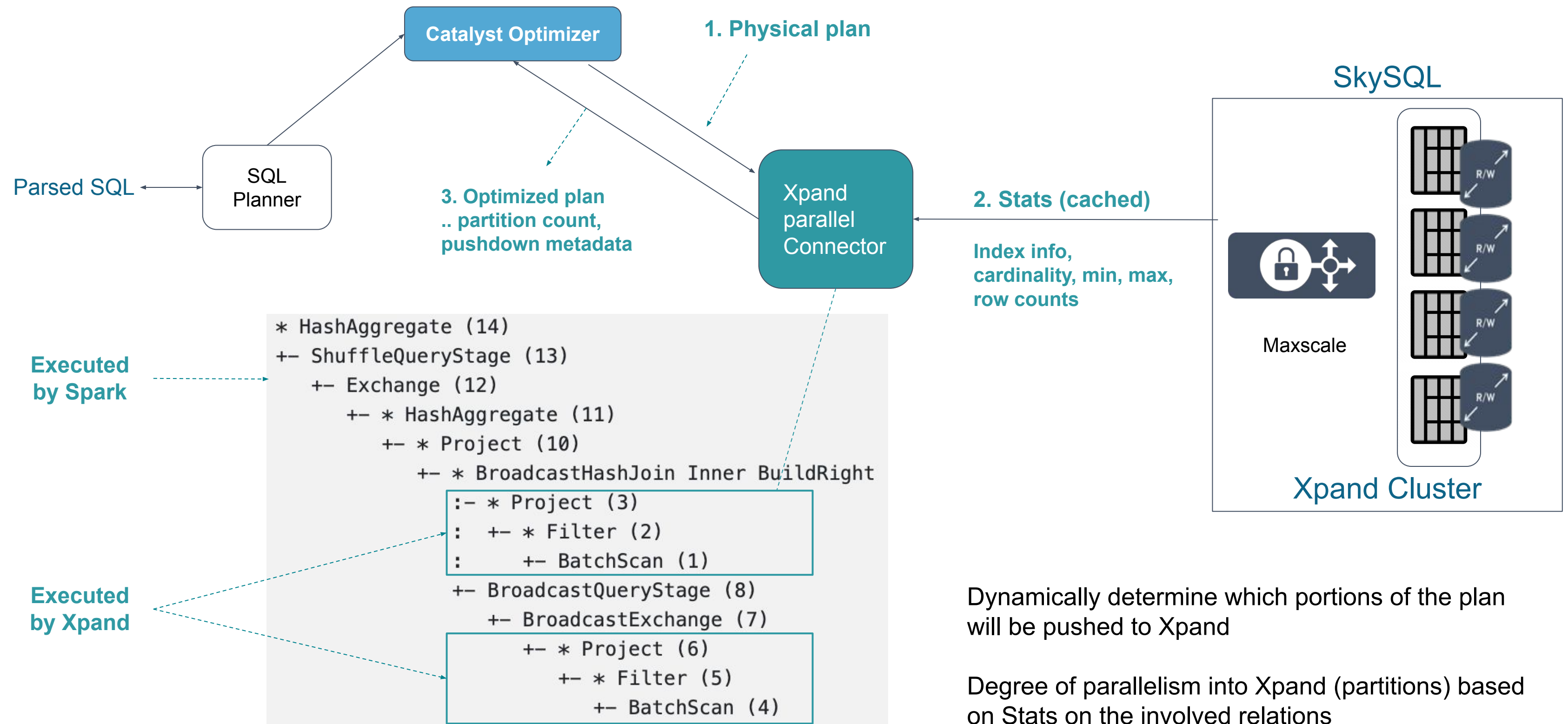
SERVERLESS ANALYTICS - PHASE I ARCHITECTURE



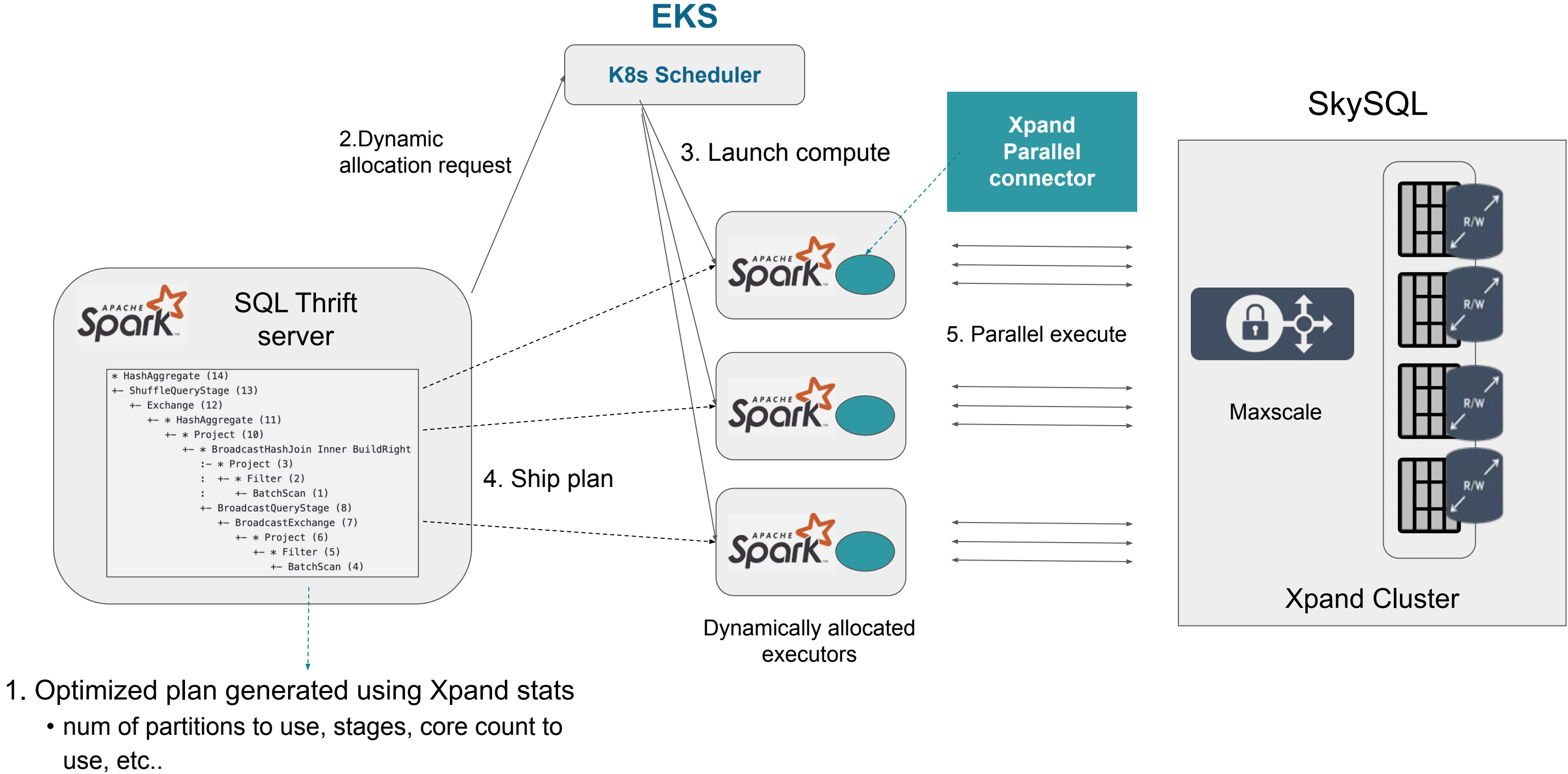
Key Point - Only the Hive server(SQL endpoint) and Notebook server run in steady state using cheap 2 core pods. All Analytic processors (executors) are dynamically allocated by K8s



PLANNING ADAPTIVE PARALLELISM, QUERY PUSHDOWN



ADAPTIVE PARALLELISM, SMART QUERY PUSHDOWN



DEMOS

- Zero-ETL Analytics on data stored in SkySQL Xpand
- Zero-ETL, Federated join on data stored in SkyQL Xpand and SkySQL MariaDB
- Federated join query between SkySQL Xpand and S3 data
- Fast, parallel ingest data from S3 into SkySQL Xpand

USE CASE PATTERNS

- **Use Case 1 - Data migration - Loading, offloading, data pipeline**
 - Move data from MariaDB to Xpand, your data lake(cloud storage/Hadoop) to Xpand/MariaDB, Snapshot DB to S3, etc.
- **Use Case 2 - Cheap but high performance, scalable Analytics for MariaDB**
 - We automatically discover the SkySQL Database catalog ... just load your data into a SkySQL database instance. Run your periodic reports without a separate Analytic DB.



USE CASE PATTERNS

- **Use Case 3 - Advanced operational Analytics for Xpand POCs**
 - Offload analytical queries so cluster isn't overloaded for OLTP
- **Use Case 4 - Interactive Analytics over data lakes (flat, semi-structured and any format)**
 - Support easy access to any Cloud storage, Hadoop like EMR.
- **Use Case 5 - Federated Analytics ... Blend data in Xpand with data lakes**

NEXT STEPS

Check out these sources to learn more about MariaDB

- **OpenWorks sessions to watch live**
 - Using MariaDB ColumnStore with Power BI for Visualization and Reporting
 - Panel: Best Practices for Migrating Your On-Premises MariaDB Deployment to SkySQL
 - SkySQL vs. AWS RDS vs. GCP Cloud SQL
 - How Hughes Achieved Scalability and High Availability for Their IoT Smart Plug with SkySQL + Xpand
- **OpenWorks sessions to watch OnDemand**
 - SkySQL - The Open, Unified and Most Productive Cloud Database for Modern Applications
 - Better Together: MariaDB SkySQL running on Google Cloud Platform
- **Try SkySQL for free**
 - [Try the full SkySQL service with a \\$500 credit, including ticketed support](#)



THANK YOU



OPENWORKS

BE UNSTOPPABLE