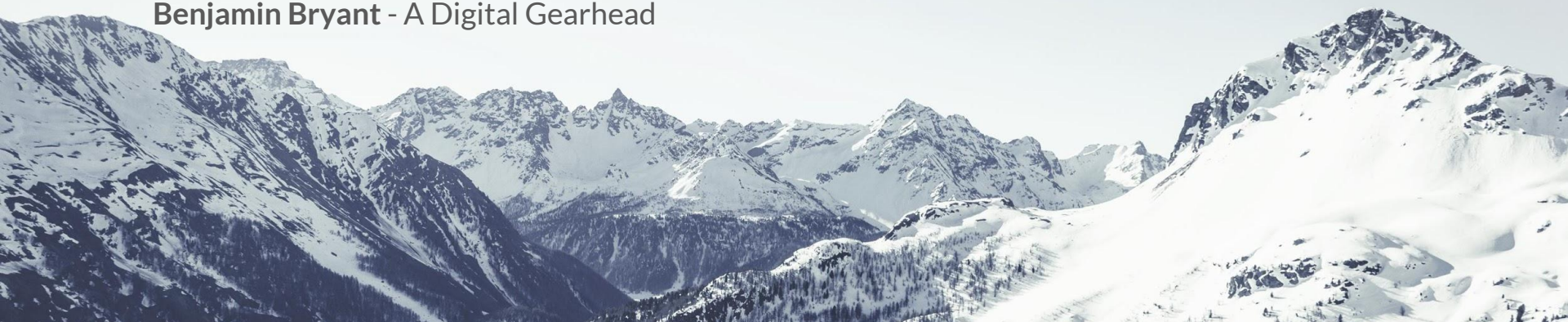
Snappy Data Exploration

From Zero to Hero Using AMD Ryzen Threadripper, JSON Documents, and Massively Parallel SQL++

Benjamin Bryant - A Digital Gearhead



Background

We want to qualify more reference data. But, COVID-19 crisis is causing additional cost pressures.

We now have the proliferation of impressive multi-threaded hardware capabilities at even the consumer/user level.

HOWEVER - can technical users and teams conveniently leverage these hardware capabilities to *accelerate* data exploration objectives in the current business climate? If so, how?

I will demonstrate and approach using AMD Ryzen Threadripper and Couchbase Analytics.

Problem - find more good data

1. Facilitate the exploration of larger unfamiliar datasets (10's to 100's of millions of records) with speed and agility.

2. Take advantage of assets we own, but without locking us into them. We expect to transition to a more operational cost model (cloud) when the economy recovers.

3. Minimize time to first meaningful result (<1 hour). Use familiar languages and techniques.

4. Minimize researcher effort that is not directly related to the assessment of data value.

Project objective

Establish a data exploration discipline to quickly qualify new valuable sources of intelligence - using cloud-friendly technology but deployed on idle assets - requiring minimal user instruction and with minimal operational friction.


AMD Ryzen
Threadripper and
Couchbase

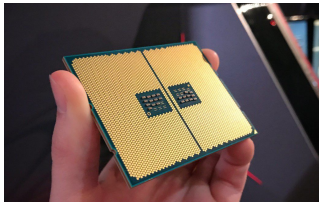
Massive Parallelism

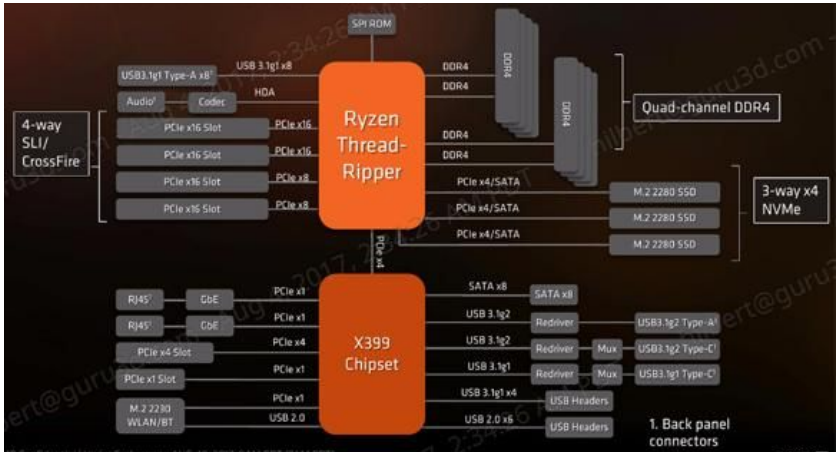
00

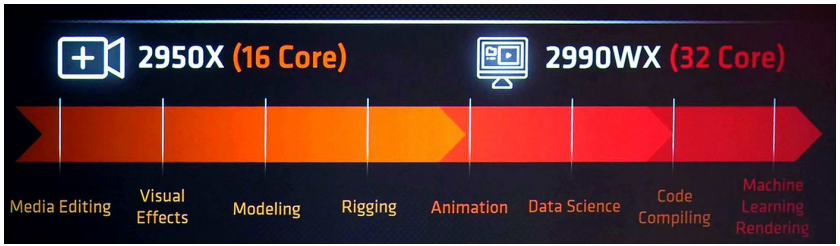
AMD's Ryzen Threadripper platforms provide next generation multi-core densities combined with massive memory bandwidth.

Project Implications:

- Less need to cluster.
- More processing power near the user.
- Flexibility to iterate faster.
- Lowest cost/CPU cycle density.



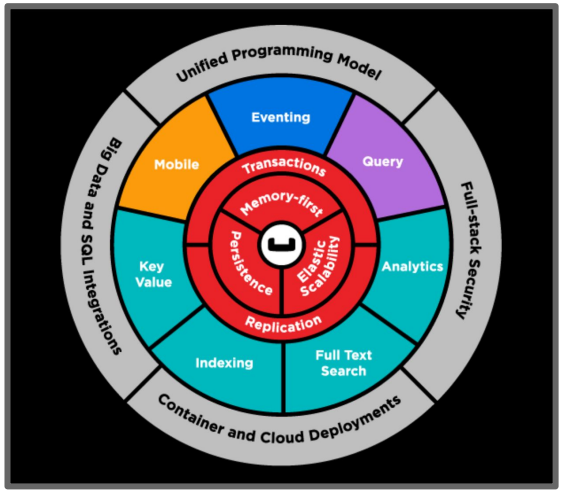




NoSQL Technology

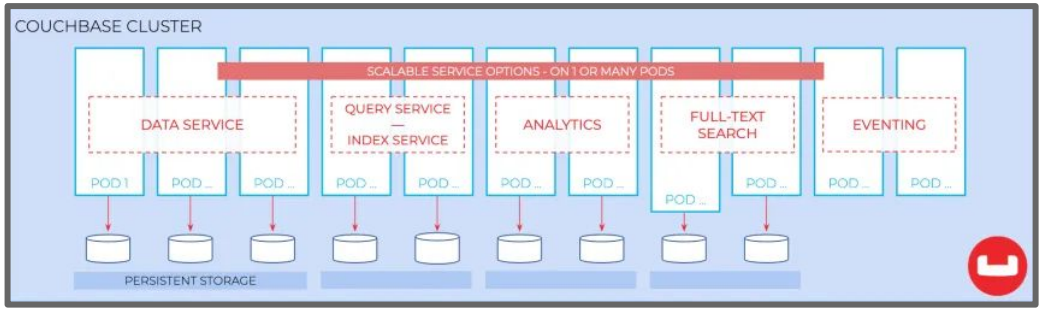
01

Couchbase' NoSQL core is perfect for supporting workloads requiring massive IO with flexibility and minimal friction.



Project Implications:

- Non-schema-enforced. Import first, analyze second.
- Natively-hierarchical data (JSON).
- Aggressive tunable memory usage.
- Fast parallel data ingestion.
- Straightforward clustering.
- Multiple index options.
- Kubernetes Operator.
- Free to download and start using.



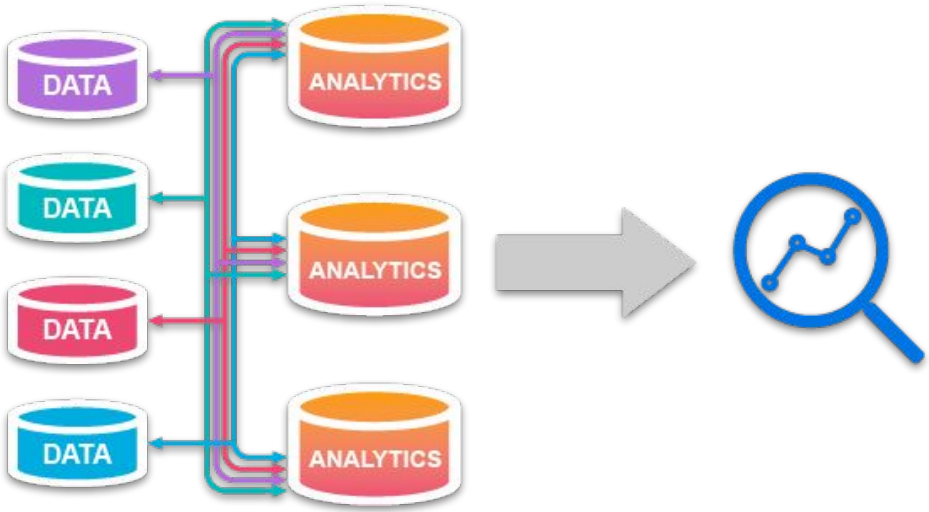
MPP Query Engine

02

The Couchbase Analytics service is a fully-integrated massively-parallel SQL++ query engine that can make maximal use of available system resources for any single query.

Project Implications:

- Simple to deploy - fully integrated into Couchbase Cluster Manager.
- SQL - minimal learning curve.
- Immediately start querying across large sets of ingested data, no indexes required.
- Sophisticated query language. Subqueries, window functions, etc.



Demo

Configure and use a local Couchbase instance for data exploration.

Steps:

1. Run Couchbase Enterprise 6.6 Docker Image
2. Configure single-node cluster for maximum analytics parallelism
3. Load and query quite a bit of data - 2 examples
4. Observe speed to first insights
   (achievable with ordinary hardware)

Demo Details

Ingest 2 datasets:

- Chicago Crimes
  https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-Present/ijzp-q8t2
  - year 2001 to present
  - 1.6GB single CSV
  - 7.2M records

- Medicaid State Drug Utilization - 10 years
  https://data.medicaid.gov/browse?category=State+Drug+Utilization&limitTo=datasets
  - Year 2011 to present
  - 5.16GB, (10) CSV's
  - 40.6M records

Using a single-node Couchbase Cluster

- Data Service
- Analytics Service

On a single workstation

- AMD Ryzen Threadripper, 16 physical cores (32 threads)
- 64GB RAM
- Dual 500GB NVME
- Docker on Linux
- Rough performance equivalent to a cluster of 4 Macbook Pros

For each dataset, ask 3 Questions:

- Sample the data; look for classifiable attributes
- Run a couple of interesting aggregations

In Conclusion

Couchbase NoSQL with its Analytics MPP query engine provides a **very fast path to first insights** into new, unfamiliar data - even on consumer hardware.

Time to install and configure

<5 min

Easy stand up
No subscription required

Time to load 40+M records

10 min

Low-Friction Ingestion
No schema required

Aggregate across all records

9 secs

Fast first results
No indexes required

** on similar hardware, see previous slide*

Next Steps...

What data would you like to explore?

Connect with me if you would like more help:
https://www.linkedin.com/in/bryantbenjamin/
https://twitter.com/bl_bryant

Thank you.

Connect with me if you would like more help:
https://www.linkedin.com/in/bryantbenjamin/
https://twitter.com/bl_bryant

Overview

The 2020 COVID-19 crisis has driven many long-standing organizations to make some adjustments in digital strategy:

1. Reduce operational costs
    a. Shift to smaller, less costly, physical assets (scrap the big airplanes, etc)
    b. Reduce workforce
2. Extract more value out of remaining assets
3. Yet, also embrace cloud-friendly technologies

At the same time, the drive to differentiate with data-driven intelligence continues:

- Catering more proactively to future customer/user concerns
- Positioning more convenient contingencies and alternatives to:
    - Changes in the customer/user environment
    - Risk factors to business operations
- Proactively identifying new business opportunities

Therefore, we are witnessing an aggressive expansion into a sustained exploration for qualified sources of intelligence data, but with the need to do <u>even faster while at reduced cost</u> .

Agenda

Overview

Problems to solve

Project objective

The Couchbase Approach

NoSQL Technology

MPP Query Engine

Demo

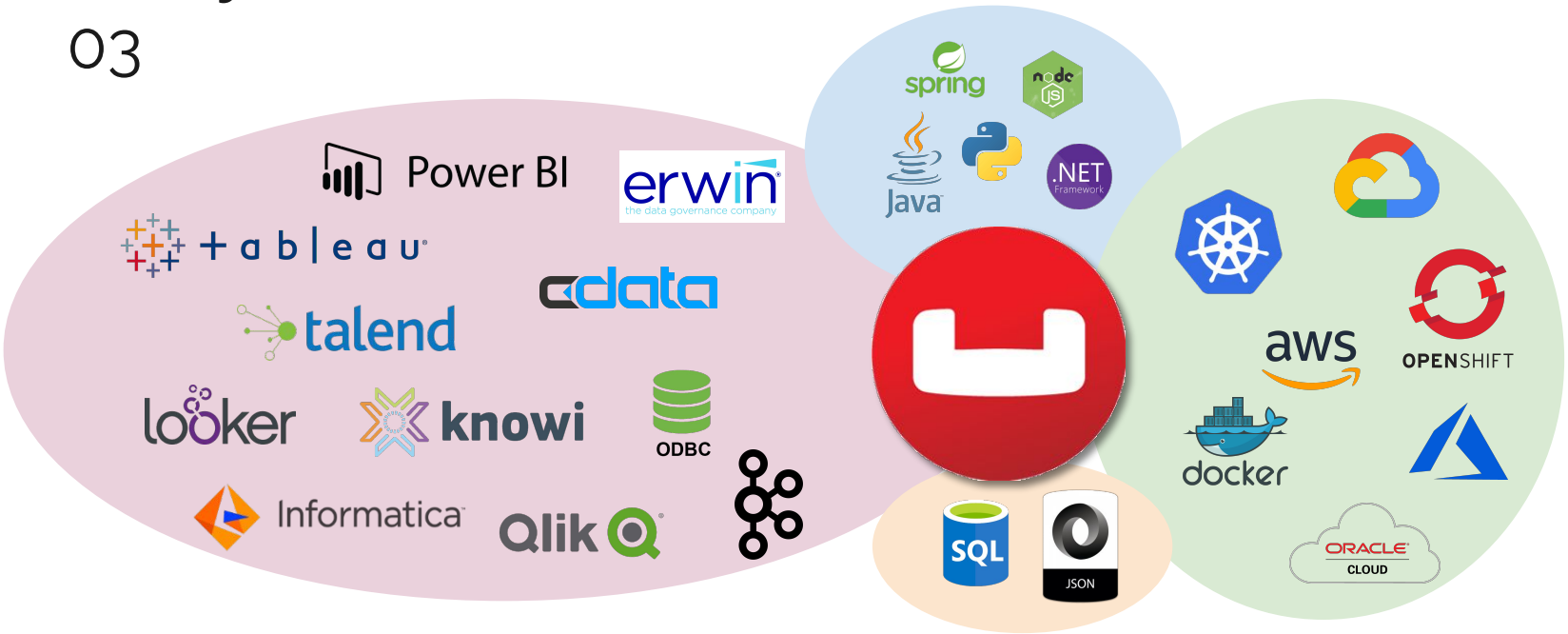
In Conclusion

Next Steps

Ecosystem

03

Couchbase is well-positioned within the enterprise data and cloud ecosystems.



Vs the Alternatives

04

Couchbase brings speed and flexibility to data analysis.

Software: Datastax, Teradata, Exadata, Hyve, Vertica		**DBaaS:** Redshift, Big Query, SQL Warehouse, Oracle Cloud, Snowflake
Complex to acquire, deploy, configure and use.	First results with speed; Couchbase Cloud Analytics coming soon	Lot's of capacity, but often slow without paying much more for performance.
Expensive to try, difficult to move.	Simple to try. Onsite and cloud options compatible and interchangeable.	Cloud lock-in