Open Source & Data Virtualization

By Antoci Rosario – 2020-11-30 (Module 4 - BigDive)

~$whoami

- Open Source Evangelist
- Italian Linux Society - Board Director
- IT Infrastructure Specialist at HPE CDS

After this lesson you should:

Open Source:

- Comprehend the benefits of open source movement (product solutions, community benefits, career development etc..)

- Data (Acquisition / Preparation / Virtualization / Visualization / Exploration) Pipeline!
- Solid knowledge of the infrastructure level, containerization and dataflow,
- Importance of data virtualization in substitution/support of general ETL solutions.

Part I - Open Source – Recap of a 22 years

What is open source?

Officially born in 1998, has its roots on Free Software movement founded in the 80s by Richard Matthew Stallman.

It’s basically started when the Linux Project was sustainable (1992/1993)

Open Source movement is guided by OSI (Open source Initiative)

Part I - Open Source - Community - Why it matters?

Open Source model is only sustainable with less constraints as possible,so the best entities who can really provide assistance, further development, documentation, beta testing and many other form of support are communities.

It’s useful for individuals to get into their local communities to learn and share open source knowledge.

At first there were Linux User Groups, now there are meetups for any kind of technology trends and even companies follows the “community model”.

Part I - Open Source – Enterprise – Why it matters?

- Different approaches to businesses: from the economic model to the organization perspective
- Cooperation and collaboration, less competition on technology discoveries
- Less static, more creative
- No forms of lock-in
- Ultimate Open Source Report by RedHat Inc.

Part I - Open Source – Enterprise – Business model

- Professional services

Open-source software can also be commercialized from selling services, such as training, technical support, or consulting, rather than the software itself.

- Software as a Service

  This kind of business is discouraged by FSF, but it's spreading since the Internet Era.

- Crowdsourcing, several core library, sometimes forgotten live by that (Ex. GNU gpg, sponsored by Facebook and Google).

Once upon a time in… datacenter:

RDBMS: single/multiple instance of same type

BI: simple

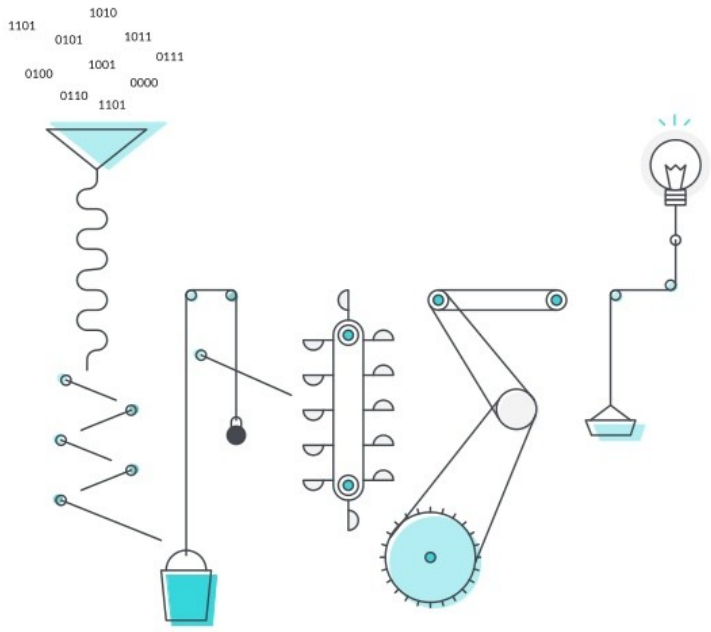
Standard reports

Ready, steady, go!

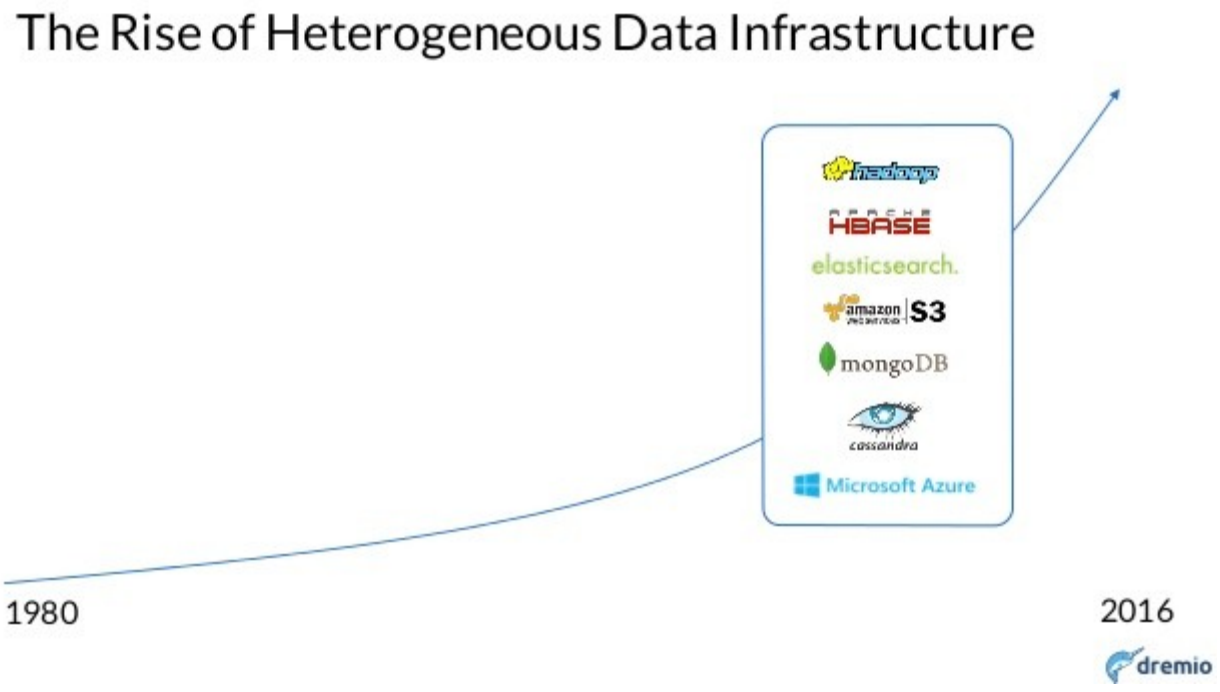
Nowadays...

Part II – Data Virtualization

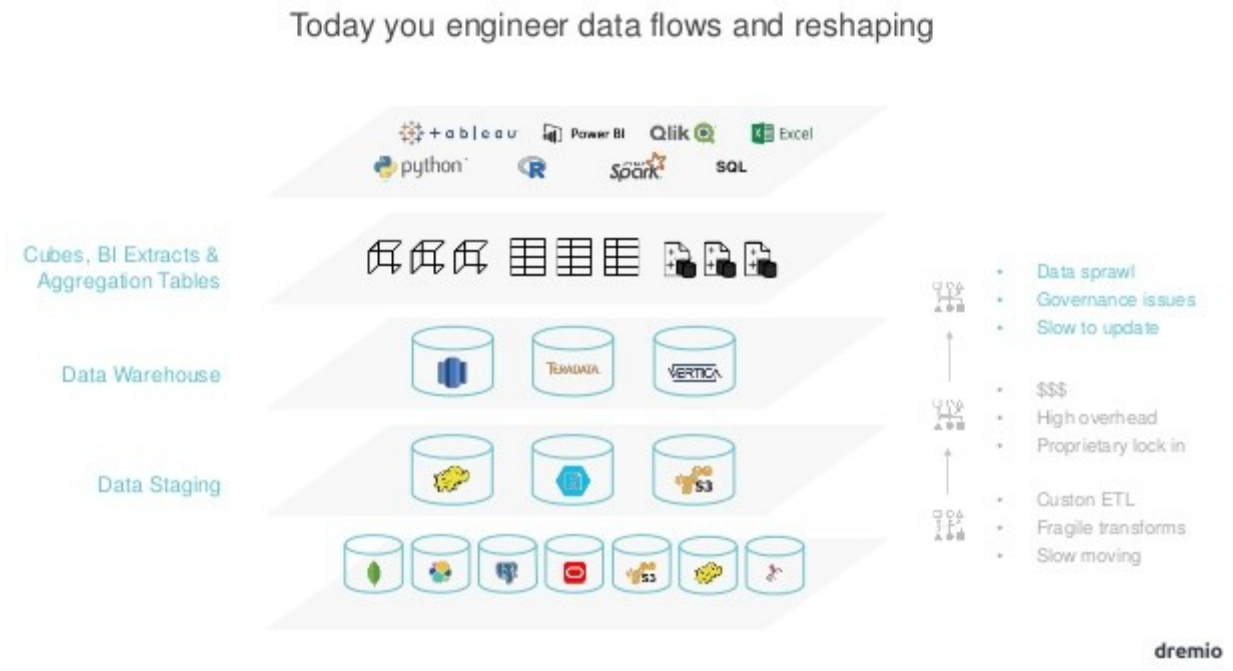
Analytics on modern data is incredibly hard



Part II – Data Virtualization



Part II – Data Virtualization

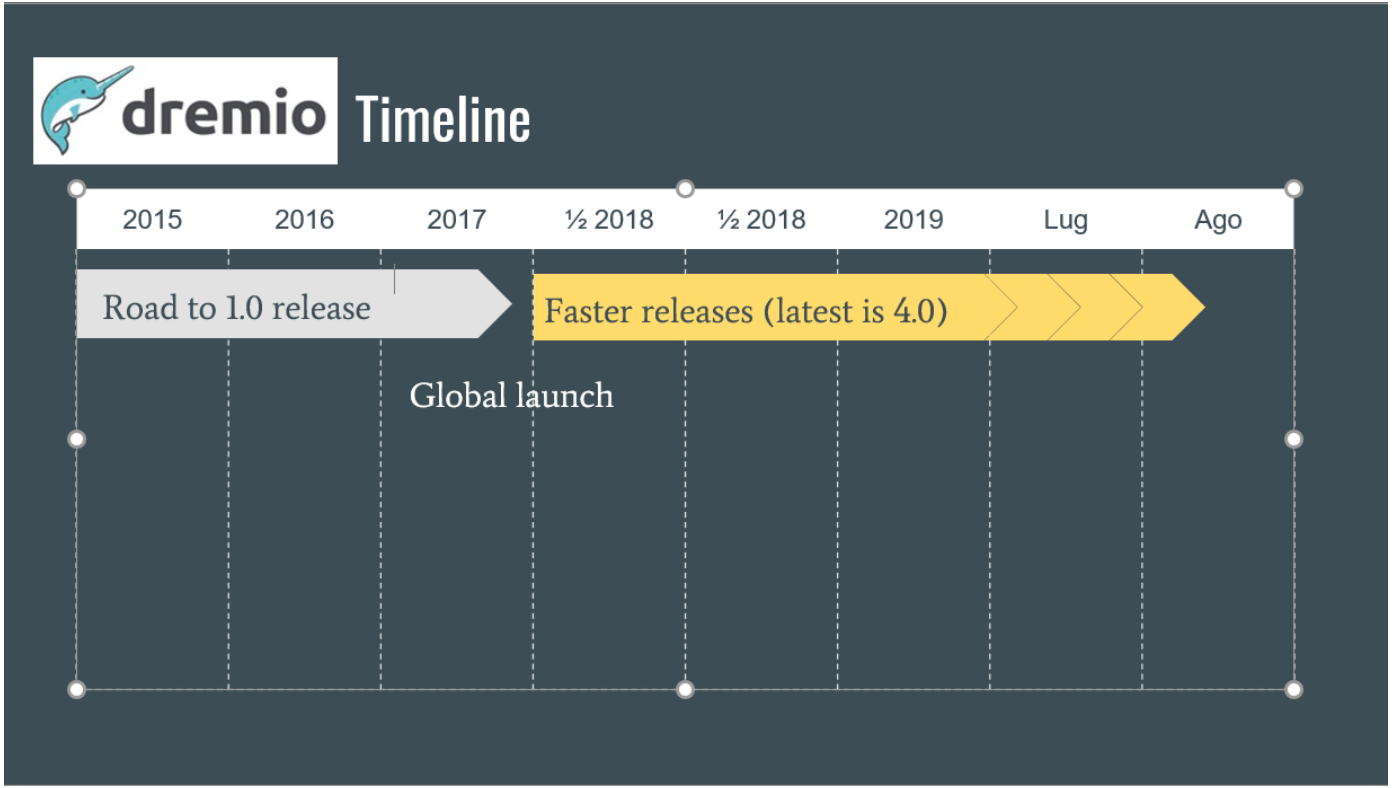


Part II – Data Virtualization with Dremio

How Dremio can help?

(since 2015)

Part II – Data Virtualization with Dremio



Part II – Data Virtualization

APACHE ARROW

What is it?

- Apache Arrow is an open source project that enables columnar in-memory data processing and interchange, optimized for memory and CPU (Intel)
- Used by an incredibly amount of projects Python and R above all.
- Columnar on memory

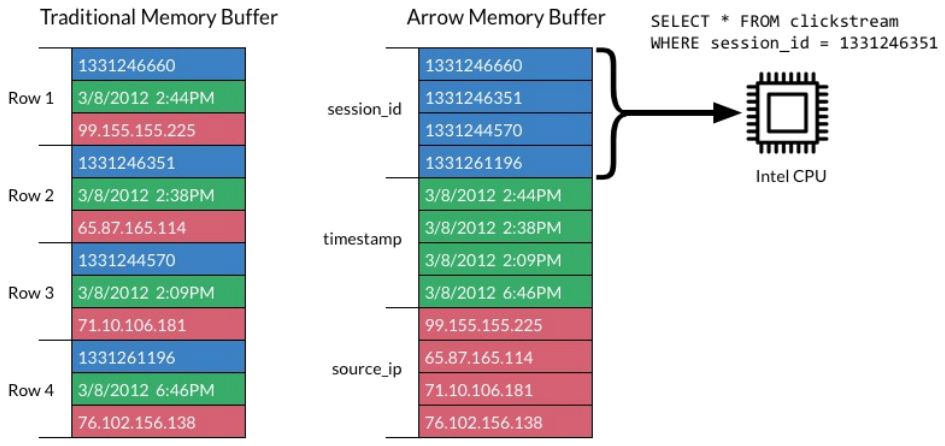
Dremio is the first execution engine built on Apache Arrow

What does it means?

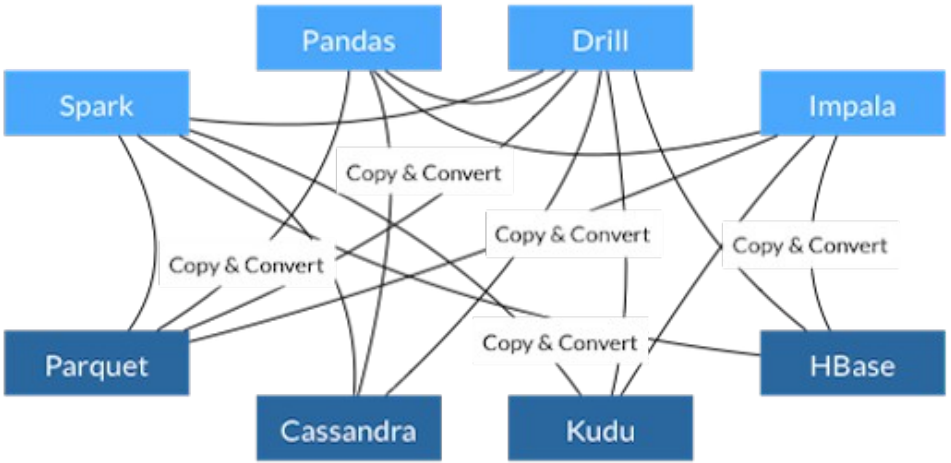
- **SIMD Paradigm**
- **Really fast reads (over 10 GB/s)**
- **Data in-memory is maintained off-heap in the Arrow format avoiding slowness.**

Part II – Data Virtualization

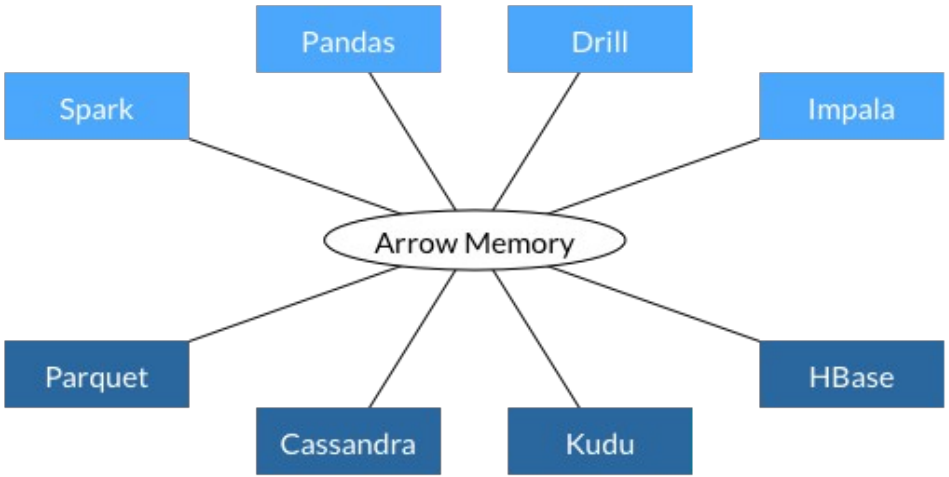
	session_id	timestamp	source_ip
Row 1	1331246660	3/8/2012 2:44PM	99.155.155.225
Row 2	1331246351	3/8/2012 2:38PM	65.87.165.114
Row 3	1331244570	3/8/2012 2:09PM	71.10.106.181
Row 4	1331261196	3/8/2012 6:46PM	76.102.156.138



Part II – Data Virtualization – Interactions before



Part II – Data Virtualization – Interactions after



Apache Parquet & Apache ORC (Columnar on Disk)

What are they?

Apache Parquet and Apache ORC are open source projects that enables columnar data storage.

So what?

- Parquet compress efficiently
- With ORC, storing data in a columnar format lets the reader read, decompress, and process only the values that are required for the current query

APACHE calcite™ (DB or not DB?)

What is it?

Not a DB.

- Standard SQL
- Query Optimizator
- Connect to third-party data sources, browse metadata, and optimize by pushing the computation to the data

So what?

- **Useful for Dremio Data Reflections**
- **You can talk SQL to NoSQL sources.**

Part II – Data Reflections

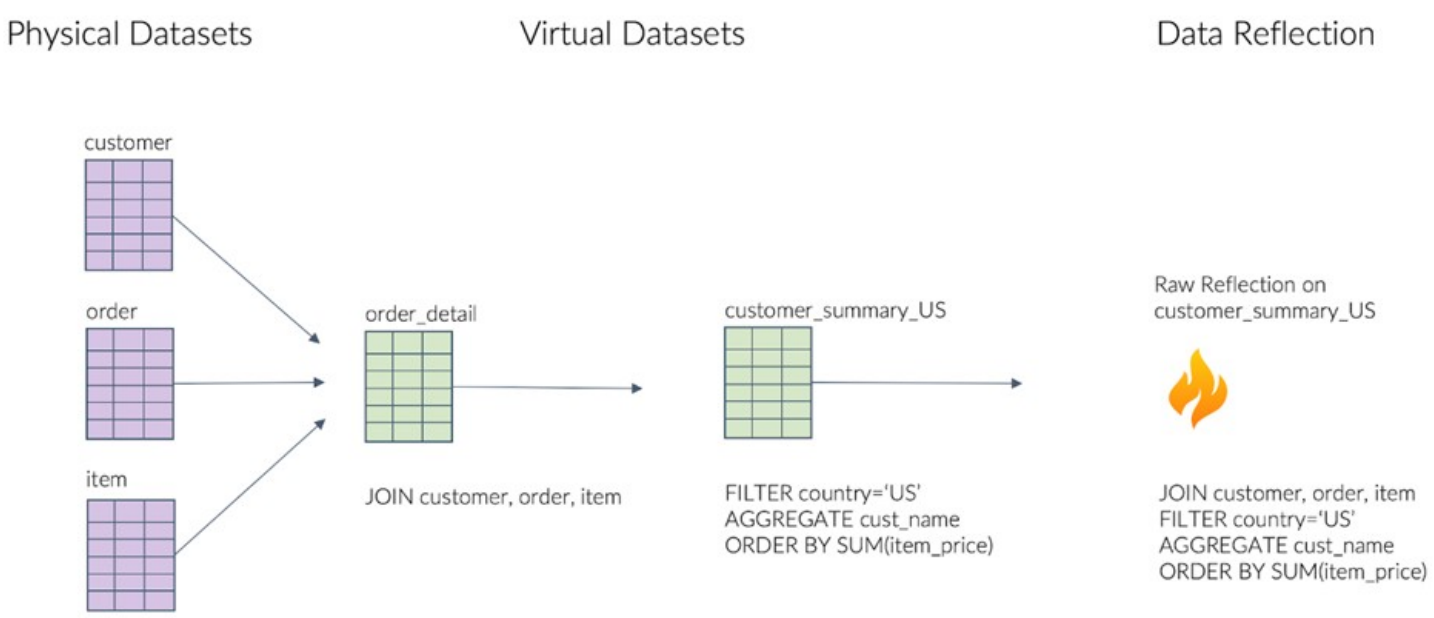
What are they?

Dremio supports two fundamental types of Data Reflections: Raw Reflections and Aggregation

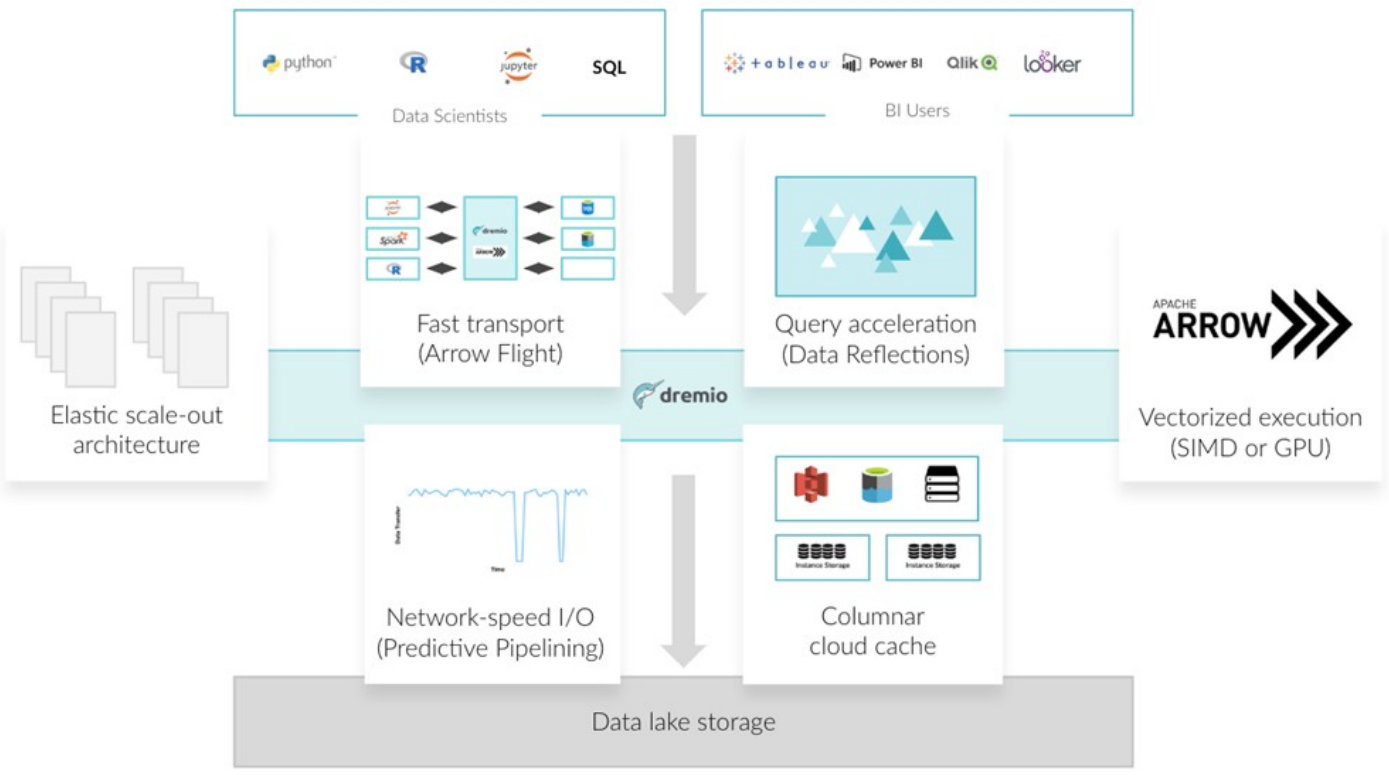
So what?

- Raw Reflections preserve row-level fidelity of the anchor dataset. A Raw Reflection includes one or morefields from the anchor dataset, and is sorted and partitioned by specific fields. You can use Raw Reflectionsto perform a number of optimizations
- Aggregation Reflections maintain summary data about the anchor dataset, so are useful to reduce the dataset target (BE AWARE OF CARDINALITY!!)

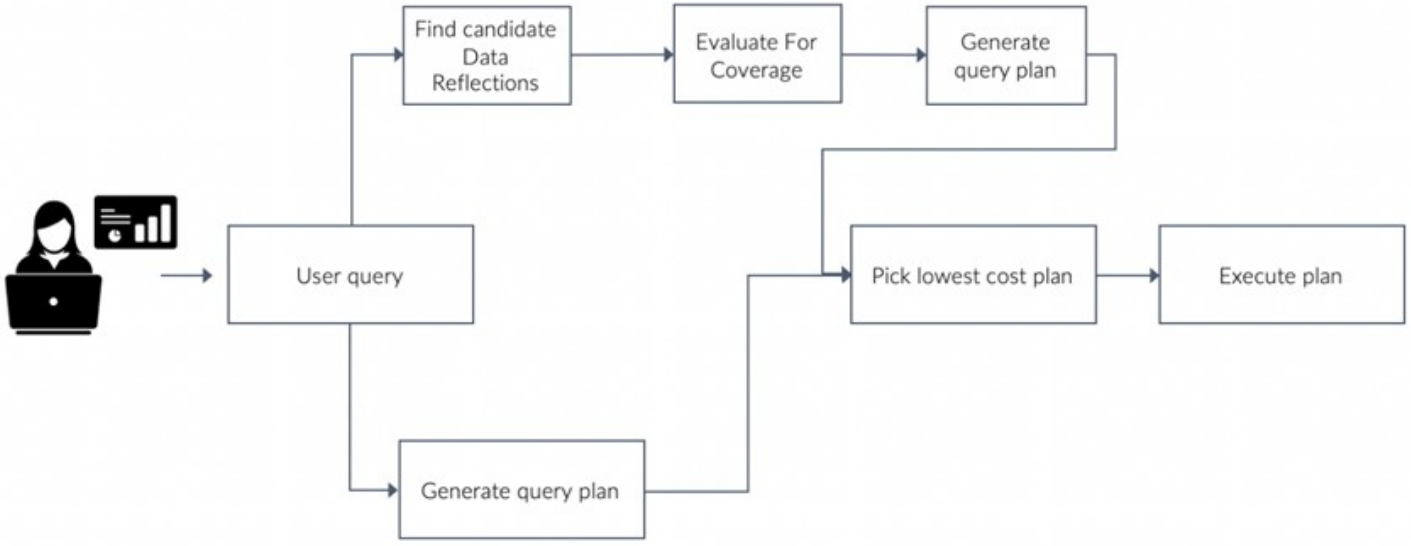
Part II – Data Reflections



Part II – Dremio utilization – Data Driven



Part II – Dremio utilization – Data Driven



Part II – Encryption and security

Dremio leverages both TLS (SSL) and Kerberos.

When connecting to a secure Hadoop cluster, Dremio communicates securely with the Hadoop services via Kerberos.

For other data sources, Dremio supports the standard wire-level encryption scheme of the source system.

Part II - Other feature C3, Gandiva...

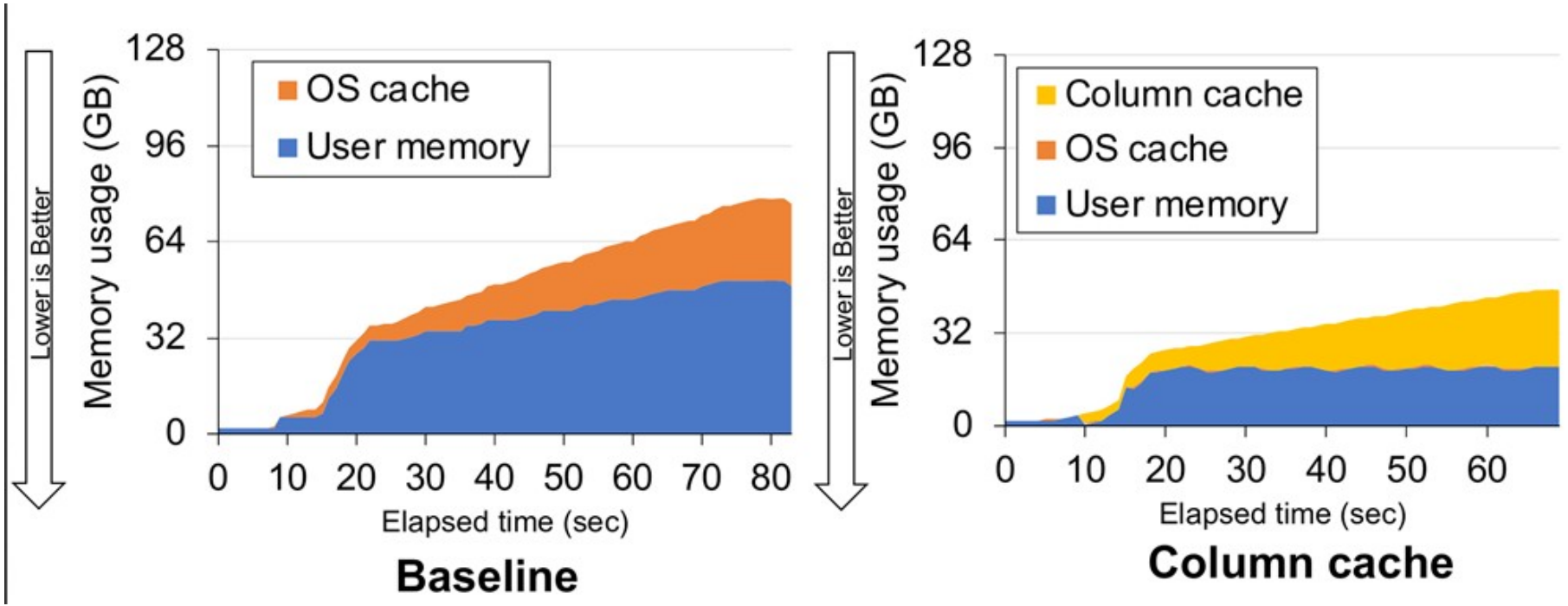
Gandiva for Apache Arrow

- Gandiva is a toolset for compiling and evaluating expressions on arrow data.

Columnar Cloud Cache

- Support for S3 and Apache Hadoop HDFS

Part II - Columnar Cloud Cache: Overview



Part II – Data Virtualization

Let’s start with a todo list…

1) Create an infrastructure to run our test, step by step.
2) Create an app for simulate data source(s)
3) Data ingestion with different methods
4) Example of Data Virtualization
5) Data Exploration and recap of the playground
6) Bonus track: Data visualization

Practical setup – Data Collection, Mock API with Python