Why data warehouses cannot support hot analytics

Gian Merlino
CTO and cofounder, Imply

June 24, 2020

What we'll discuss

- Define "temperature-tiered" analytics
- Discuss workflows surrounding "hot analytics"
- Introduce real-time data platforms as a solution
- Discuss recent performance benchmark results
- Discuss how this differentiates from popular cloud data warehouses

What is hot analytics?

Fast analytics using fresh data for all

- Fast - sub-second query response
- Fresh - streaming, real-time (hot) data
- For all - self-service UX for business people (beyond analysts)

Not all workloads are equal

Cold

- All data is available
- Low cost
- Not performance sensitive

Warm

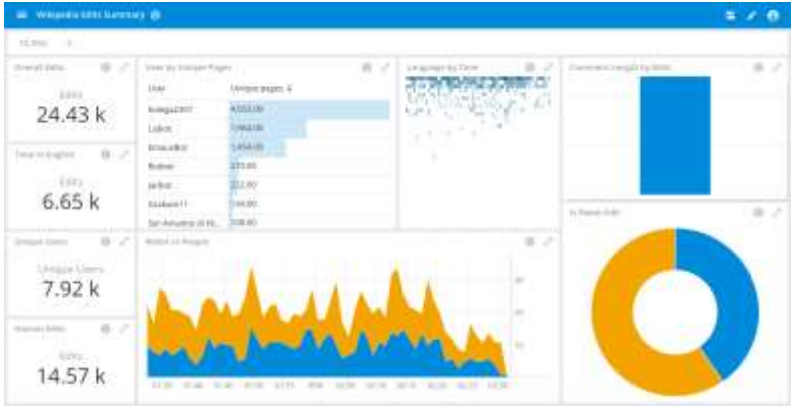
- Most data is available
- Moderate cost
- Moderate performance

Hot

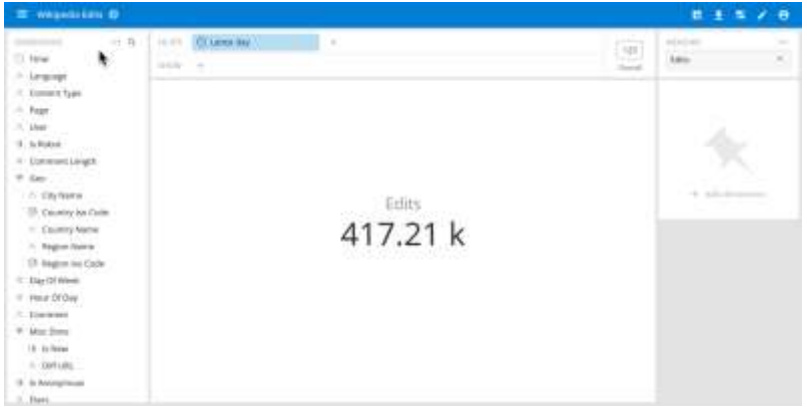
- Business-critical datasets
- Always online
- Latency extremely important

Hot data powers monitoring and exploration

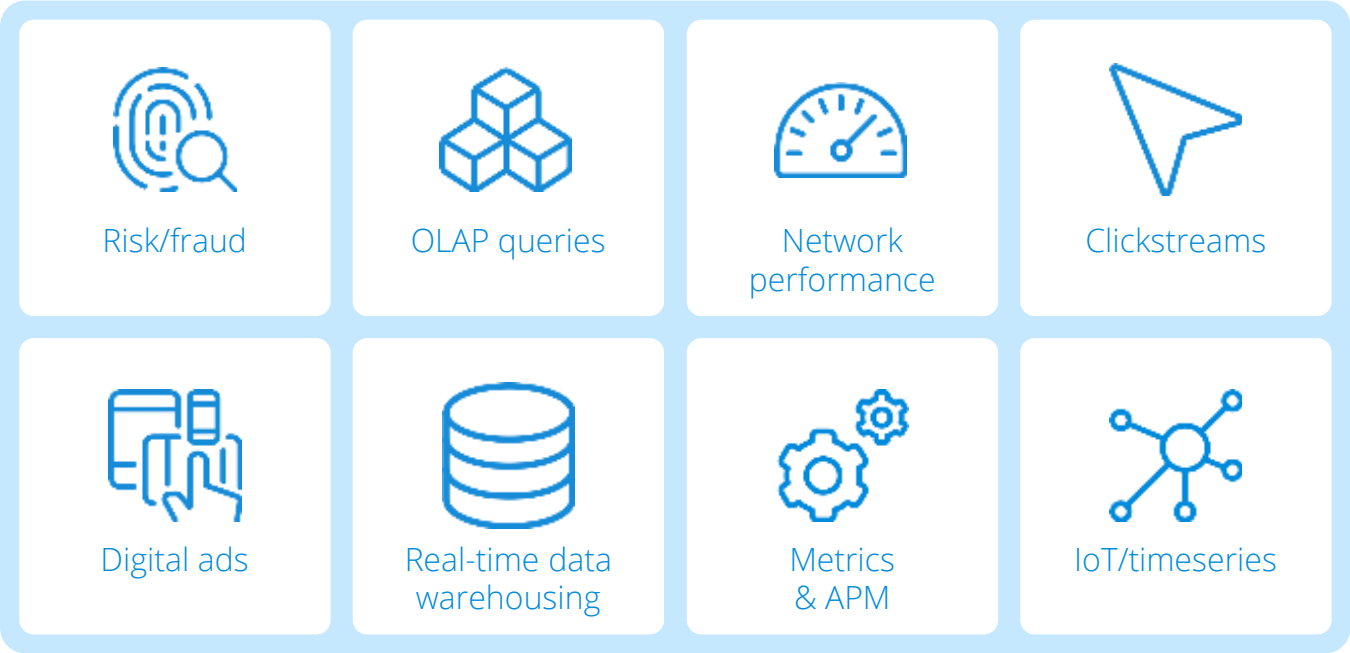
Monitor trends in dashboards



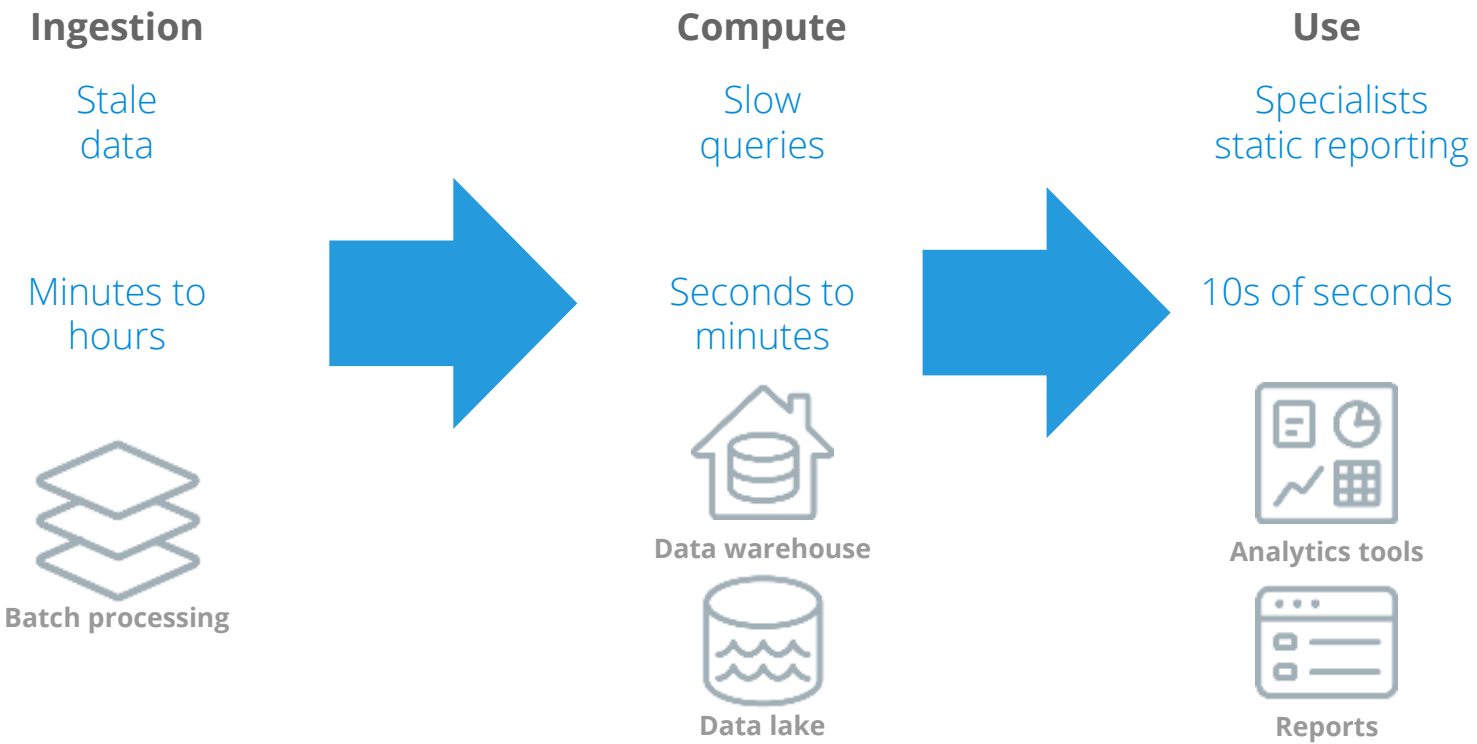
Drag-and-drop ad-hoc exploration



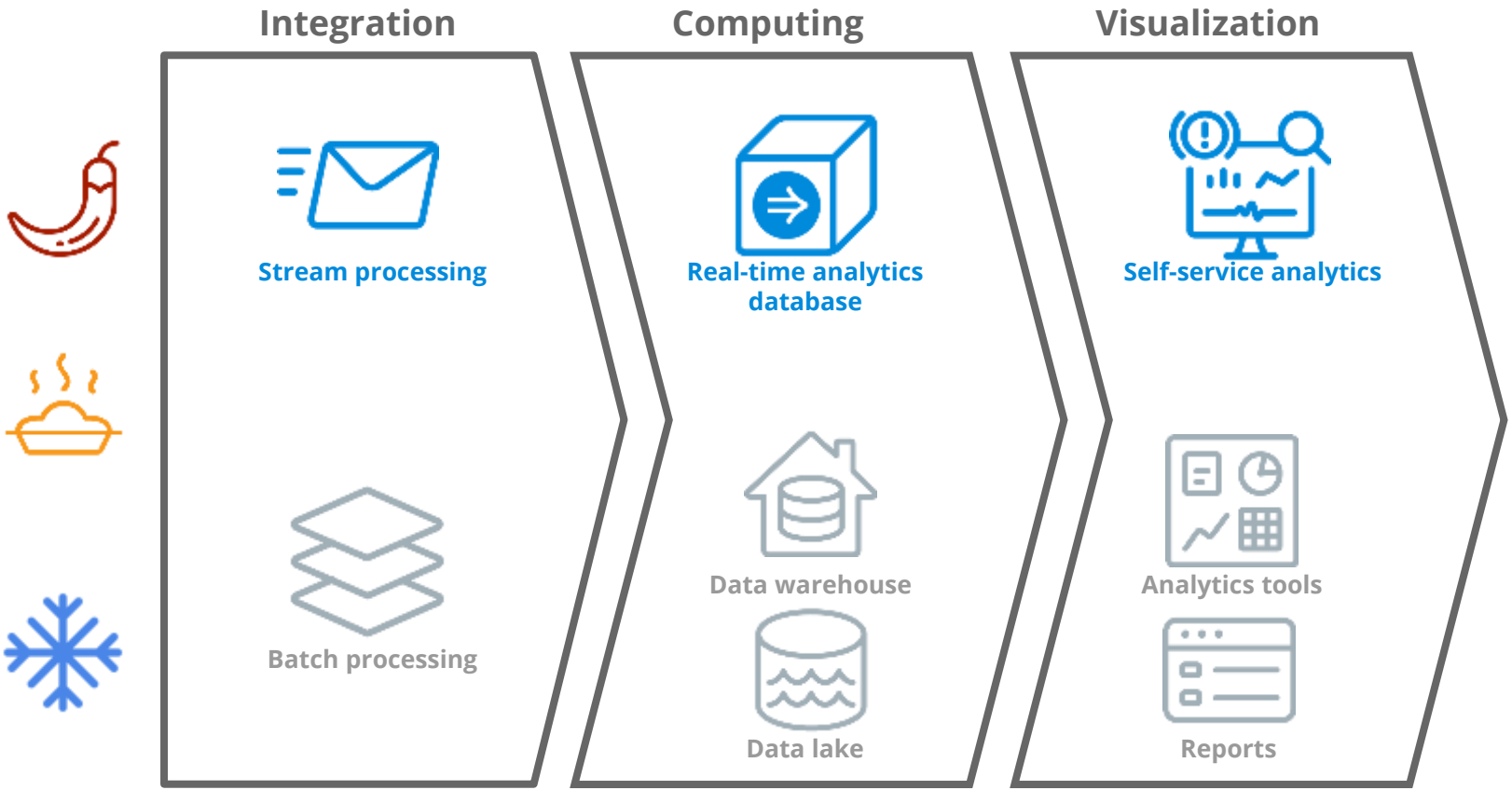
Where hot analytics is required



Data warehouses cannot support hot analytics



How to add hot analytics



A real-time data platform is required for hot analytics

Defining characteristics of a real-time data platform

- Native streaming ingestion and instant data visibility
- Vertically integrated storage, compute and visualization
- Separately scaling ingestion and querying
- Server tiering
- Query prioritization

+ Plus the standard items you expect in a modern analytics platform
  + Cloud-native
  + Elastic
  + Secure
  + Self-healing
  + Zero downtime for software upgrades

Introducing Druid

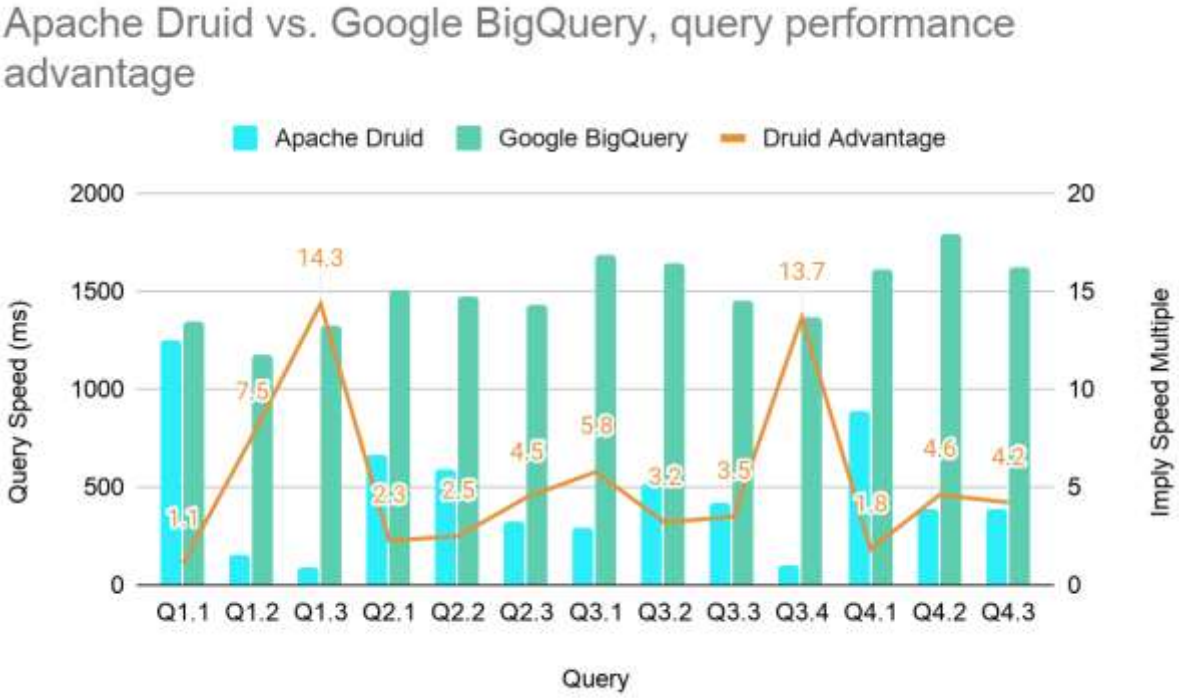
- **“high performance”:** bread-and-butter fast scan rates + ‘tricks’
- **“real-time”:** streaming ingestion, interactive query speeds
- **“analytics”:** counting, ranking, groupBy, time trend
- **“database”:** the cluster stores a copy of your data and helps you manage it

Introducing Druid

- Column oriented
- High concurrency
- Scalable to 1000+ servers
- Continuous, real-time ingest
- Query through SQL
- Target query latency sub-second to a few seconds

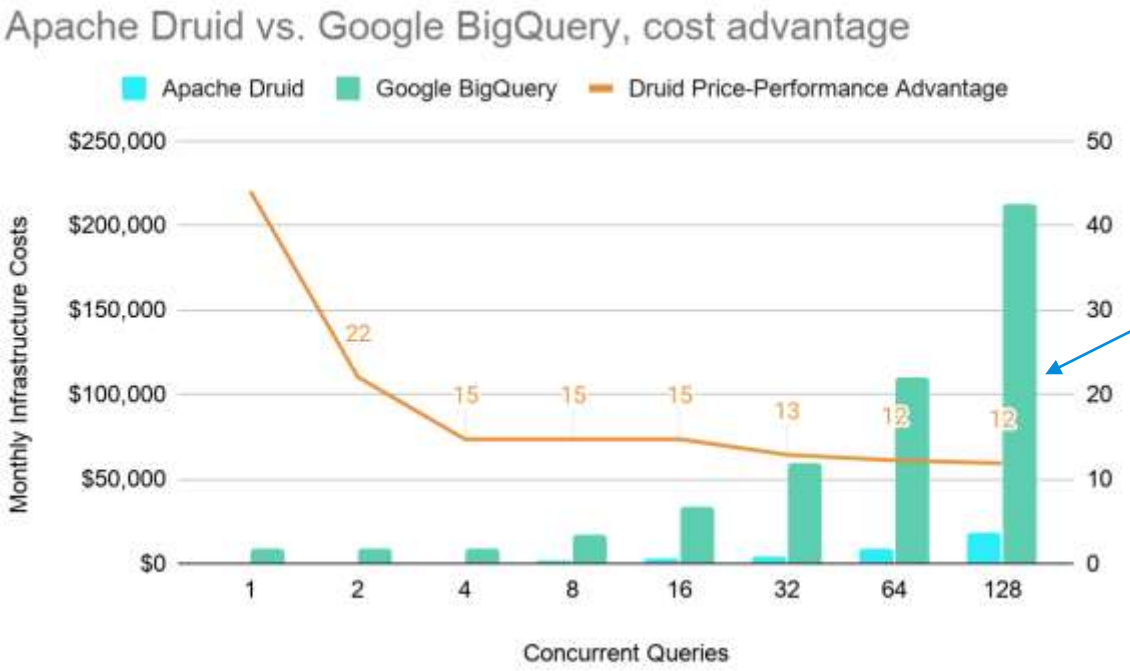
Not convinced? We ran a benchmark.

Druid vs BigQuery: 3x performance advantage



Source: Apache Druid and Google BigQuery Performance Evaluation, 2020

Druid vs BigQuery: 12x price-performance advantage



$194,600 monthly savings at high concurrency.

Source: Apache Druid and Google BigQuery Performance Evaluation, 2020

Comparing Druid and cloud data warehouses

Cloud-native w/o speed compromise

- CDWs like Snowflake & BigQuery need to retrieve data from remote storage during query execution, which slows them down.
- Druid preloads data before queries happen.

Pull-based ingestion

- Popular CDWs limit latency or throughput of real-time ingestion, if they offer it at all.
- Pull-based ingestion in Druid enables tens of millions of inserts/sec in true real-time.

Secondary indexes

- CDWs do not offer indexes beyond the partition key.
- Druid offers space-efficient, compressed secondary indexes.

Comparing Druid and cloud data warehouses

Server tiers

- CDWs give you one size that must fit all when it comes to performance and cost.
- Druid lets you control which data gets 'hot' vs. 'warm' vs. 'cold' performance.

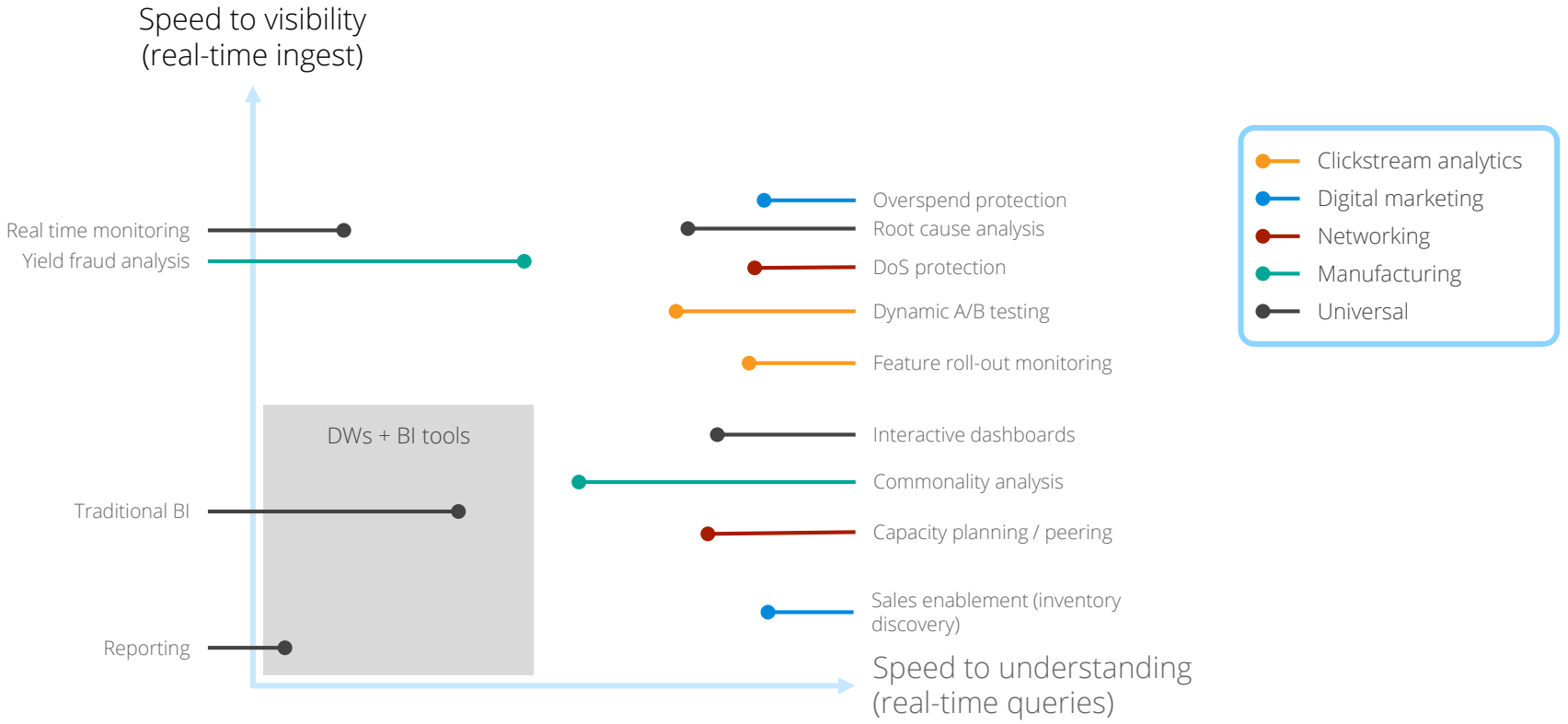
Dynamic lanes

- CDWs are not designed for running interactive applications.
- Druid uses a 'fast lane' to prioritize interactive queries over reporting queries.

Approximate algorithms

- CDWs offer some approximate algorithms, like count distinct and quantiles.
- Druid offers a wider array of approximate algorithms than any other popular database. When approximate algorithms are acceptable, they improve performance dramatically.

Our customers have unlocked many new capabilities



Questions?

Thank you!