Data Processing & Data Integration Platform

STREAM HORIZON

Big Data Analytics - Accelerated

by Threeglav © ®

THREEGLAV

stream-horizon.com

StreamHorizon & Big Data

Integrates into your Data Processing Pipeline…

- Seamlessly integrates at any point of your your data processing pipeline
- Implements generic input/output HDFS connectivity.
- Enables you to implement your own, customised input/output HDFS connectivity.
- Ability to process data from heterogeneous sources like:
  - Storm
  - Kafka
  - TCP Streams
  - Netty
  - Local File System
  - Any Other…

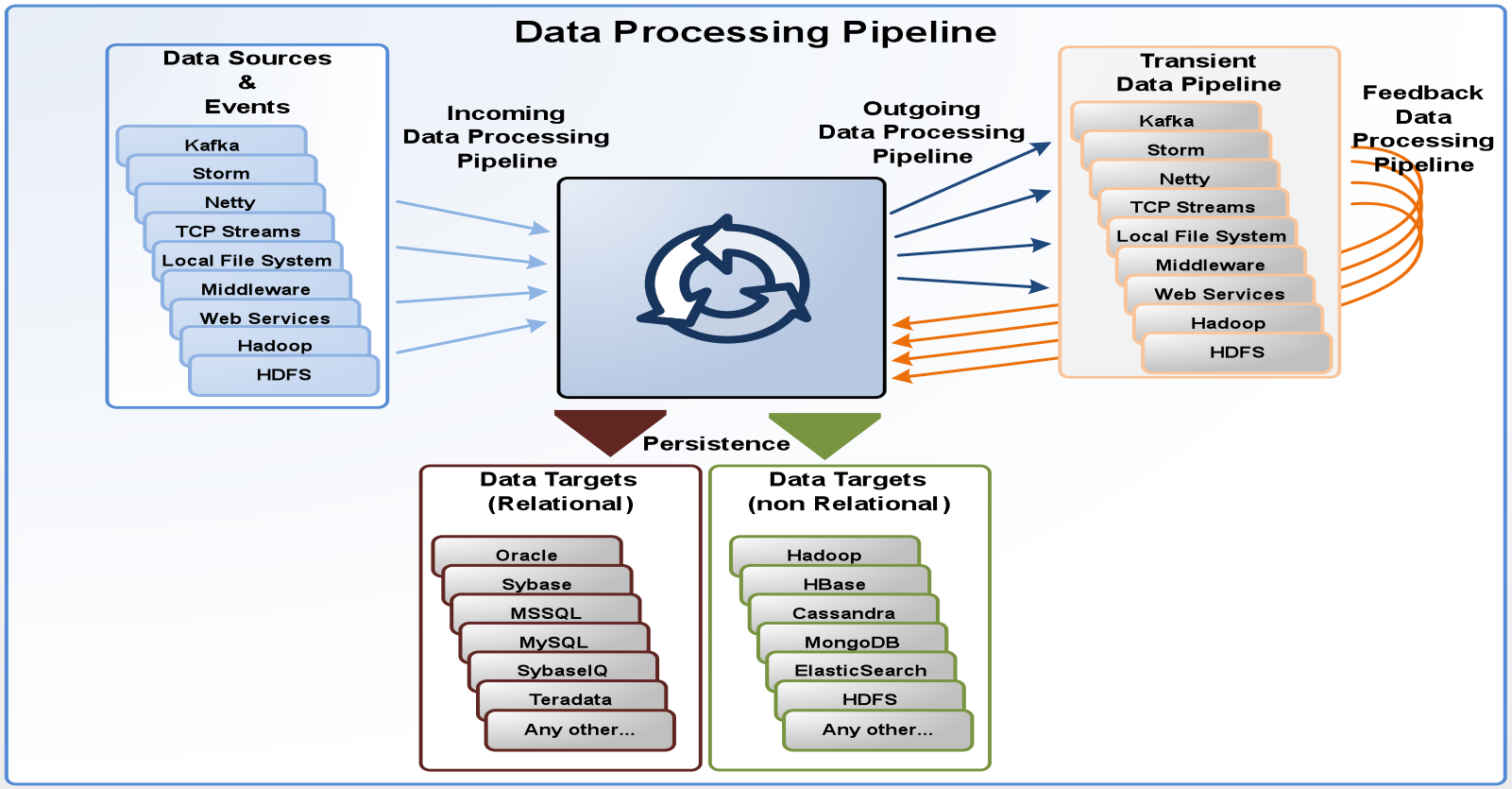
Accelerates your clients…

- Reduces Network data congestion
- Improves latency of Impala, Hive or any other Massively Parallel Processing SQL Query Engine.
  - *Impala*
  - Hive
  - HBase
  - Any Other…

And more…

- Portable Across Heterogeneous Hardware and Software Platforms
- Portable from one platform to another

StreamHorizon – Big Data Processing Pipeline



StreamHorizon - Flavours – Big Data Processing

Storm - Reactive, Fast, Real Time Processing

- Guaranteed data processing
- Guarantees no data loss
- Real-time processing
- Horizontal scalability
- Fault-tolerance
- Stateless nodes
- Open Source

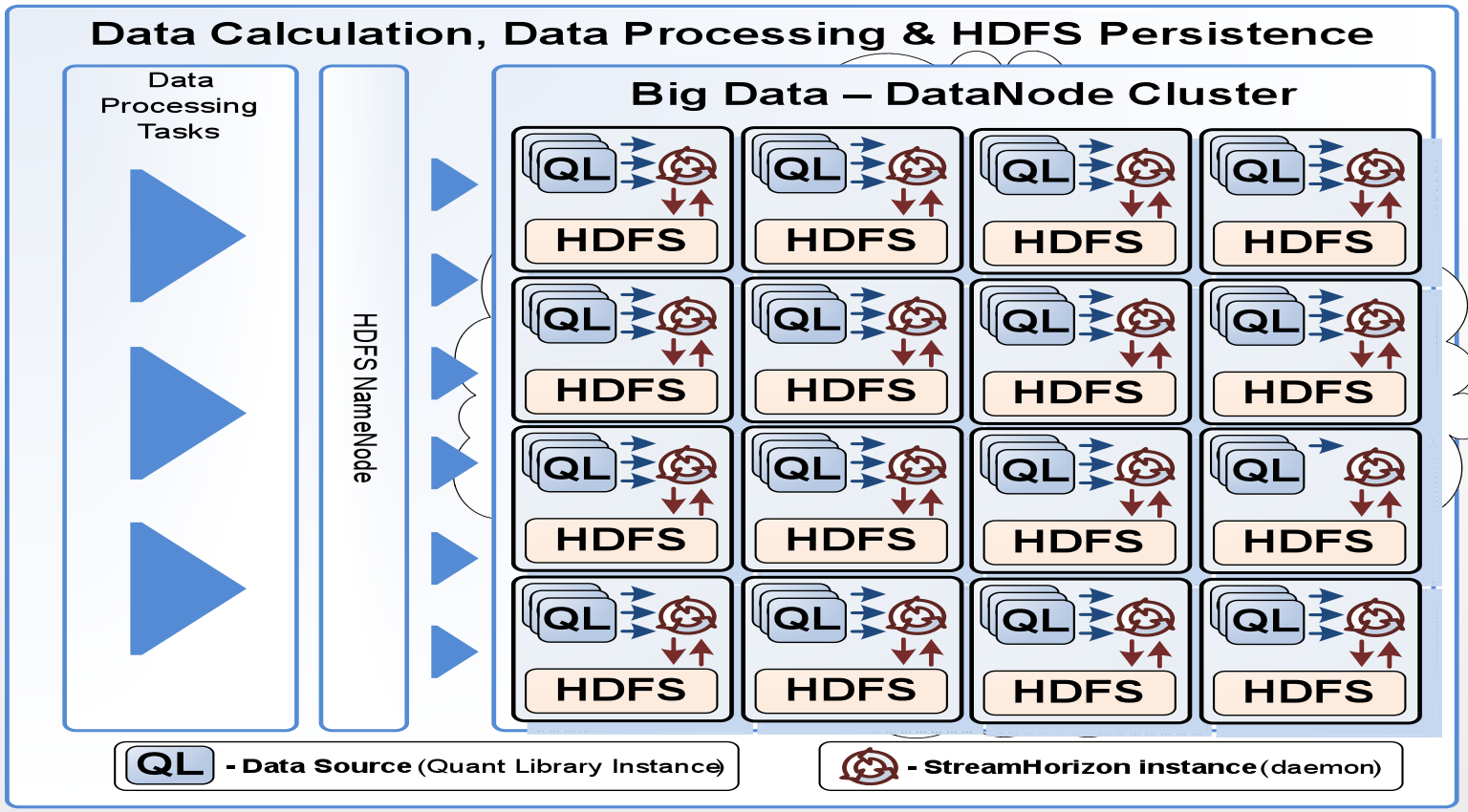
Hadoop – Big Batch Oriented Processing

- Batch processing
- Jobs runs to completion
- Stateful nodes
- Scalable
- Guarantees no data loss
- Open Source

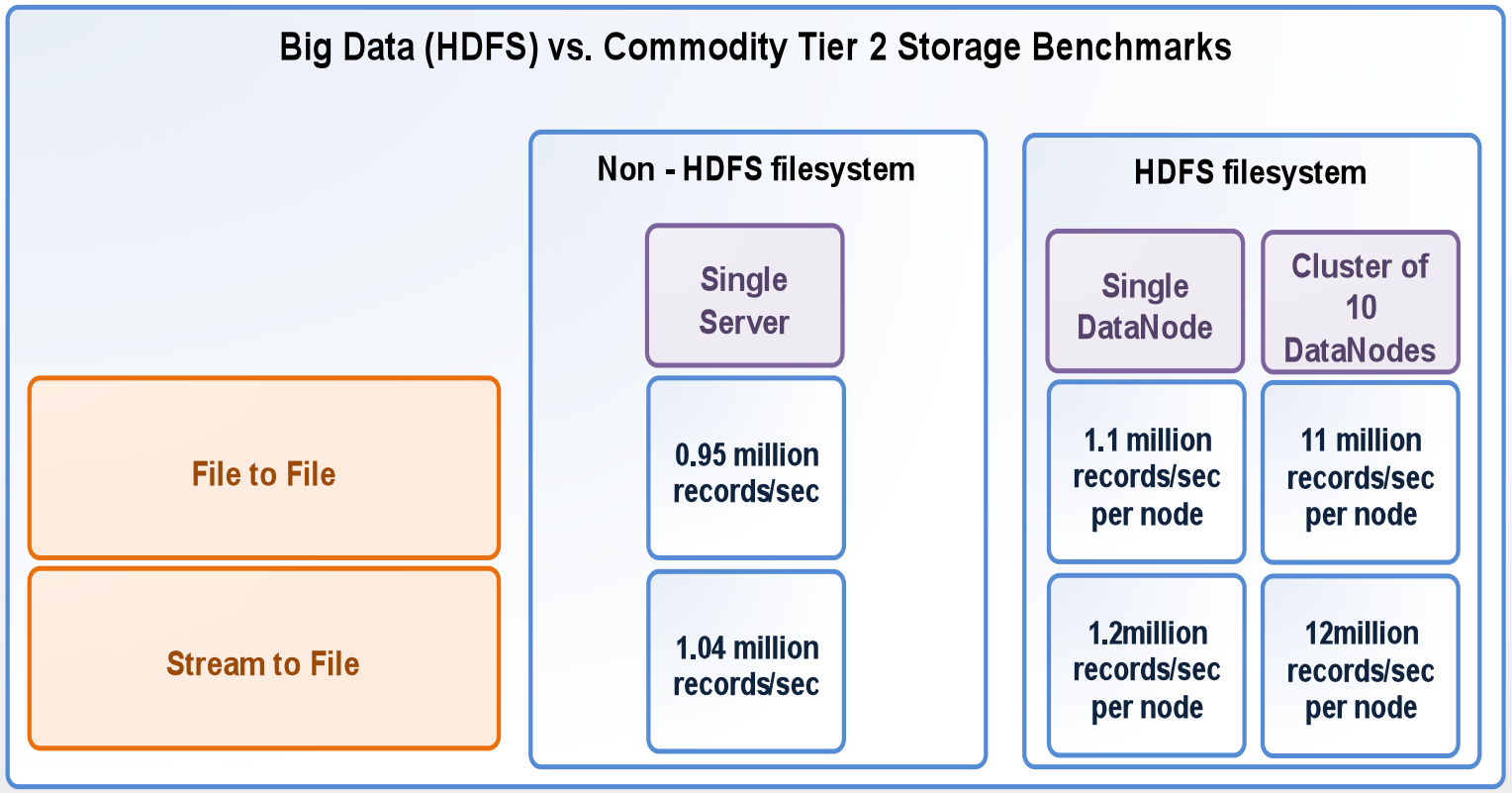
Kafka

- Designed for processing of real time activity stream data (metrics, KPI's, collections, social media streams)
- A distributed Publish-Subscribe messaging system for Big Data
- Acts as Producer, Broker, Consumer of message topics
- Persists messages (has ability to rewind)
- Initially developed by LinkedIn (current ownership of Apache)

Big Data & StreamHorizon – Data Persistence (Example: Finance Industry)



HDSF vs. Tier 2 Storage – Performance Benchmark



StreamHorizon & Big Data – Advanced Concepts

Streaming Data Aggregations – SDA

- SDA are integral part of all StreamHorizon instances (daemons)
- StreamHorizon SDA processes & aggregates data as it is processed (on the fly)
- Aggregated data is directly persisted to HDFS (or any alternative Data Target)

StreamHorizon ‘Accelerator Method’ - persisting your Big Data (faster)

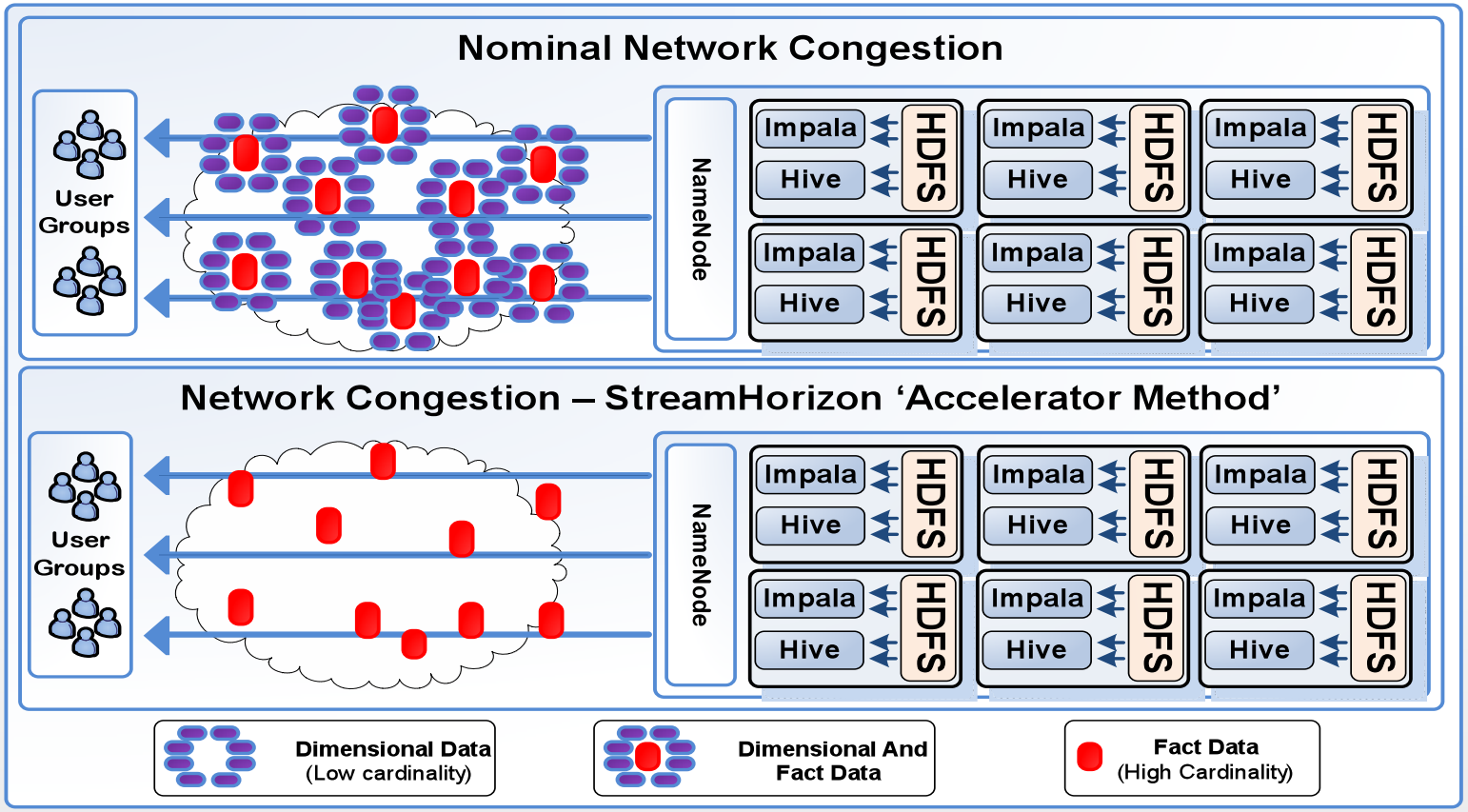
Accelerate Hadoop or/and HDFS persistence by writing less:

- StreamHorizon enables you to persist only transactional data (traditionally known as ‘Fact table data’) and omit persisting repetitive dimensional data (low cardinality data) to DataNodes across your Big Data cluster
- Client front end tool (or any end user application like Excel) simply merges Dimensional data with Fact (transactional data) brought from your Big Data cluster.

Reduced Client & Network footprint

- StreamHorizon Accelerator Method reduces Network traffic between your clients & Big Data cluster to ~10% of nominal size (due to avoidance of shipping of low cardinality data (dimensional data) via network)

Big Data & StreamHorizon – Data Retrieval (Example: Finance Industry)



StreamHorizon beneficial to Big Data filesystem (HDFS)

- StreamHorizon Accelerates Hadoop processing - MapReduce has two main disadvantages (processing is slow & inconvenient to use)
- Hive + StreamHorizon SDA outperforms Impala (Hive based reporting stack usually has higher query latency compared to Impala stack. This is significantly improved with StreamHorizon)
- Due to aggregation, single HDFS file contains even more of your business data. Benefits are:
  - Hive queries are more effective
  - Reduces number of I/O requests
  - Single I/O request executes as sequential read in comparison to default I/O footprint
  - Reduces HDFS replication latency
- Move via Network only high cardinality (Fact) rather than low cardinality (Dimensional) data. Achieve reduction of network traffic down to 10% of it's nominal value.
- StreamHorizon SDA accelerates heavy data operations like joins etc.

Client Queries - accelerated by StreamHorizon

StreamHorizon accelerates ad-hoc queries for Hive, Impala or any other MPP SQL Query Engine. This is can be achieved with:

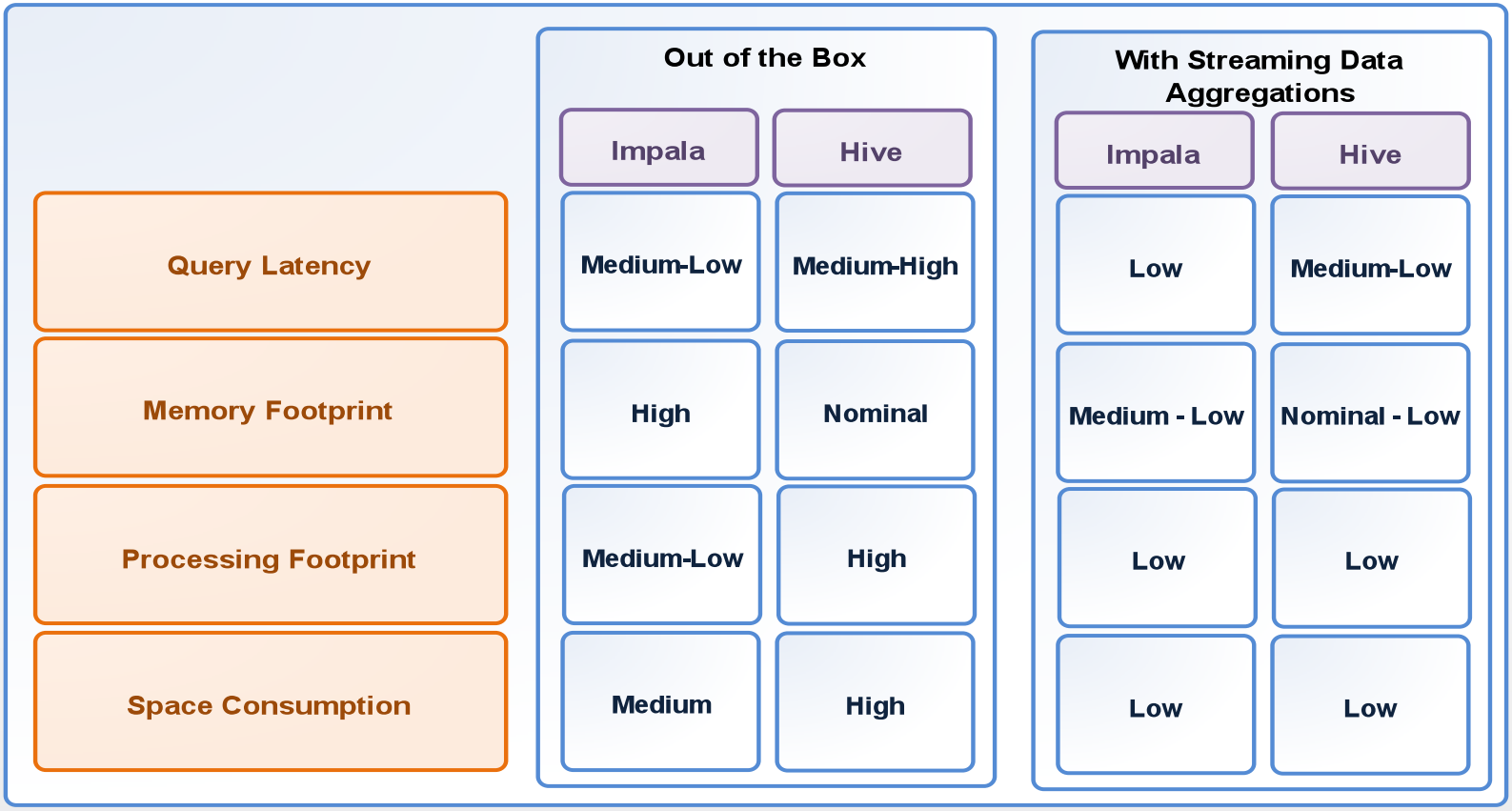
- StreamHorizon SDA (Streaming Data Aggregations)
- StreamHorizon 'Accelerated Method' data topology

StreamHorizon reduces memory pressure for Impala (or any other memory dependent data access component)

StreamHorizon reduces MapReduce processing latency for Hive (when utilizing StreamHorizon SDA)

Hive query latency reduced by order of magnitude (function of data volume reduction of your StreamHorizon SDA aggregations)

Streaming Data Aggregations – impact on Big Data Query Latency



STREAM HORIZON

Q&A