Data Processing & Data Integration Platform

STREAM HORIZON

BigData Analytics - Accelerated

by Threeglav © ®

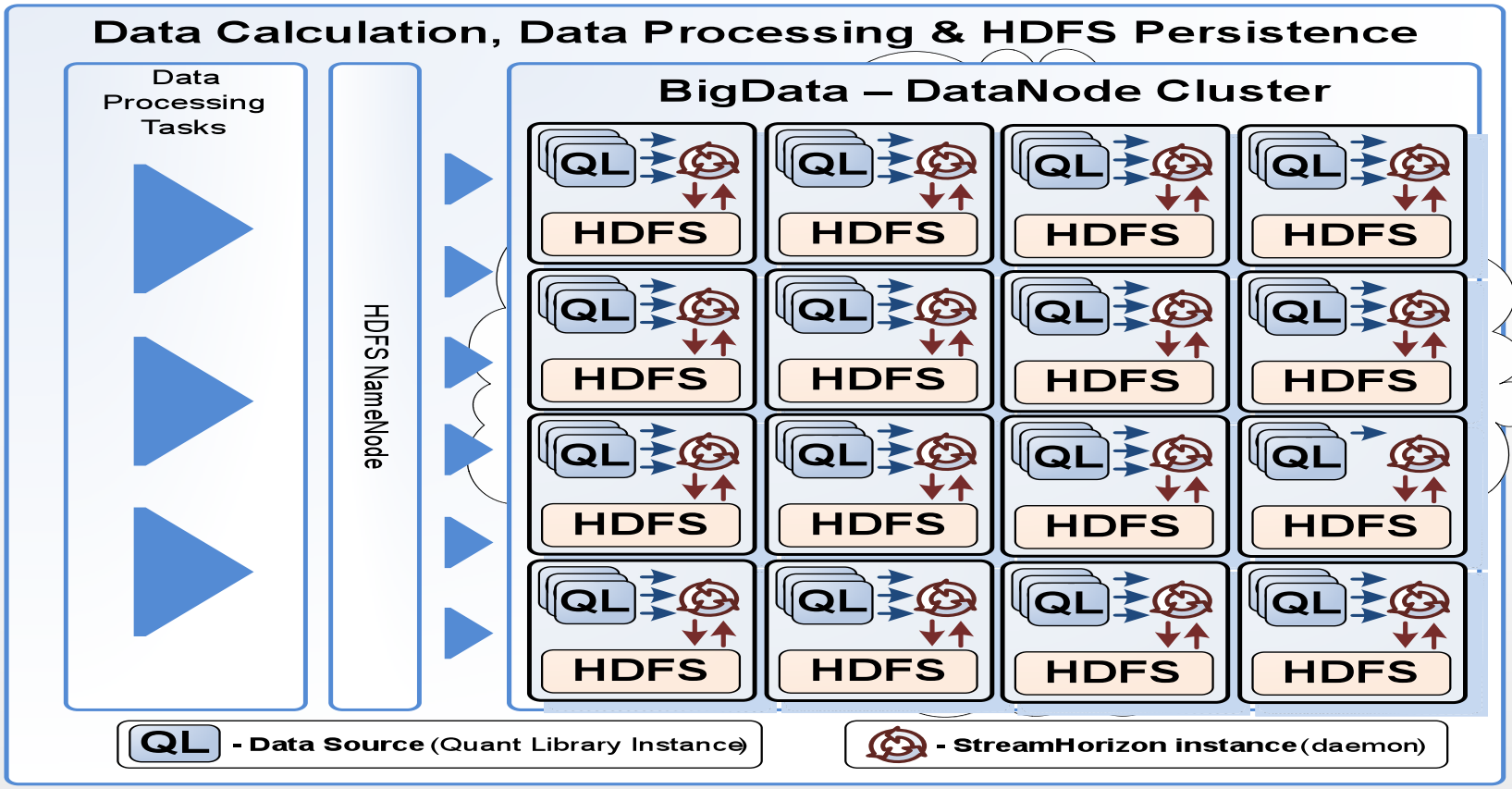
THREEGLAV

stream-horizon.com

StreamHorizon on Big Data - Vanilla deployment

- StreamHorizon instance (daemon) is deployed across all (or some) nodes of your BigData cluster (DataNodes).
- StreamHorizon implements generic input/output HDFS connectivity.
- StreamHorizon enables you to implement customised input/output HDFS connectivity.
- Ability to process data from heterogeneous sources like Hadoop, Storm, TCP Streams, Netty, Local File System etc.
- Portable Across Heterogeneous Hardware and Software Platforms
- Portable from one platform to another

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



StreamHorizon ‘Accelerator Method’ - persisting your BigData (faster)

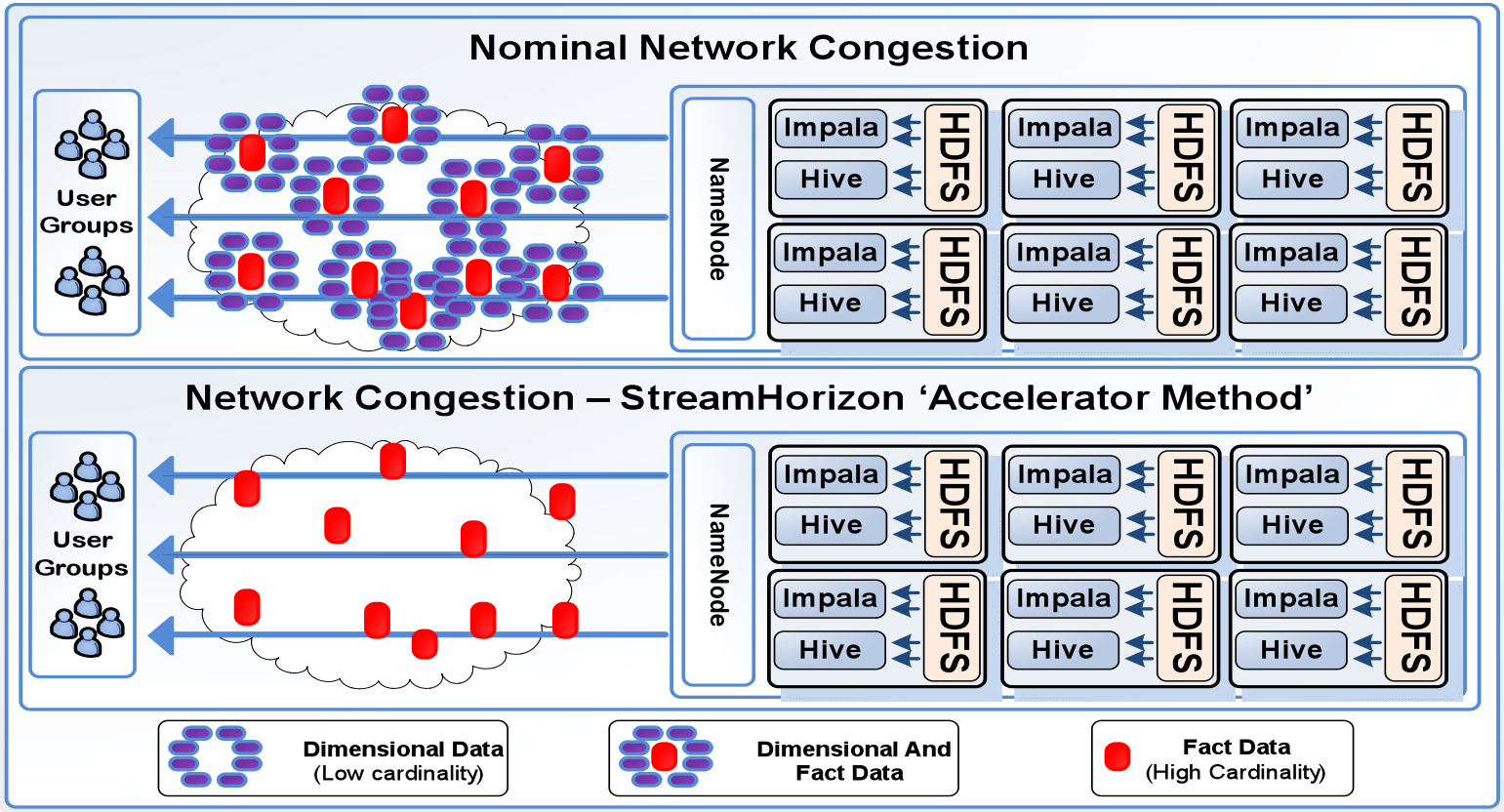
Accelerate Hadoop or/and HDFS persistence by writing less:

- 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 BigData cluster
- Client front end tool (or any end user application like Excel) simply merges Dimensional data with Fact (transactional data) brought from your BigData cluster.

Reduced Client & Network footprint

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

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

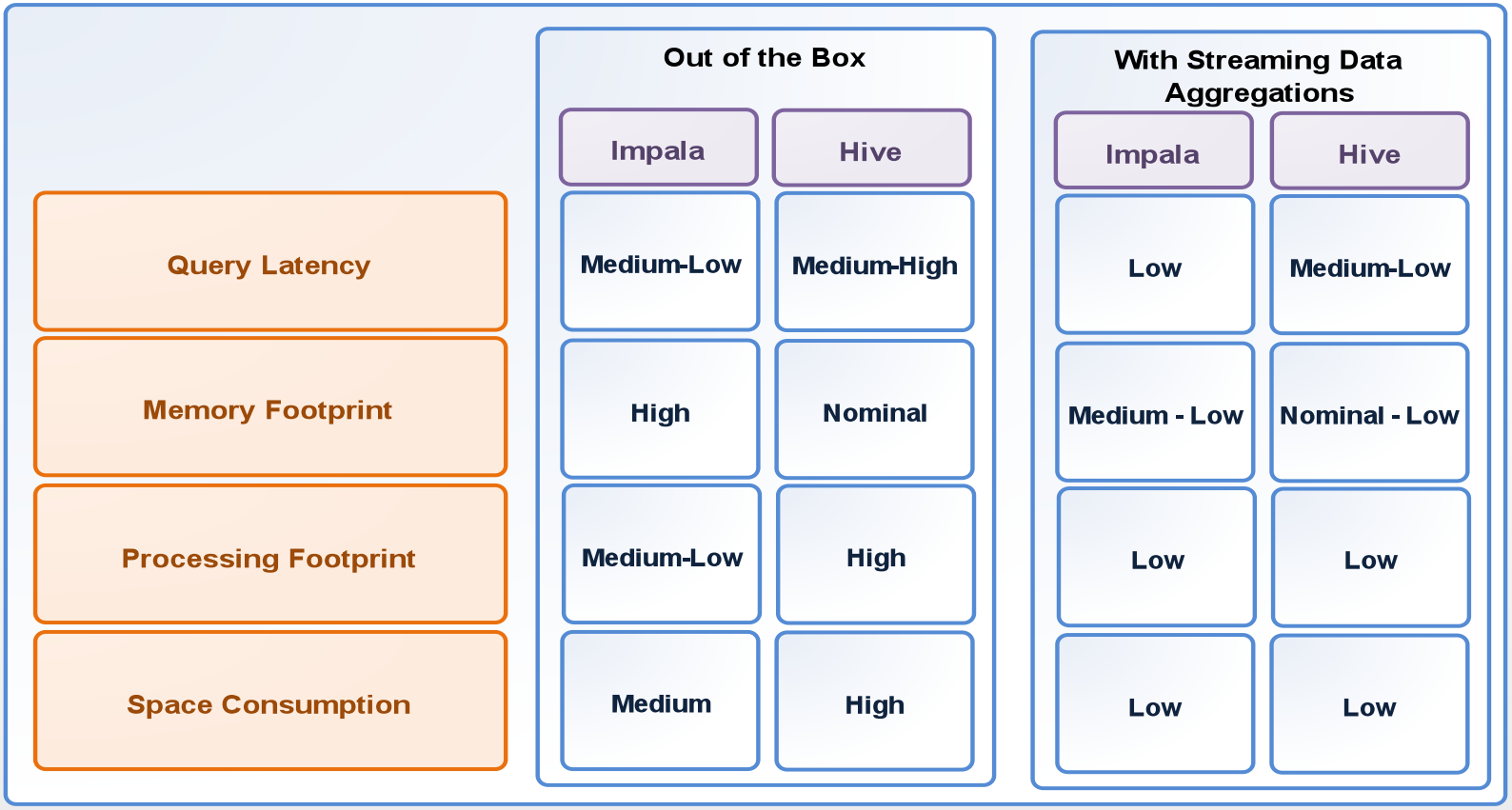


StreamHorizon & Big Data – Advanced Concepts

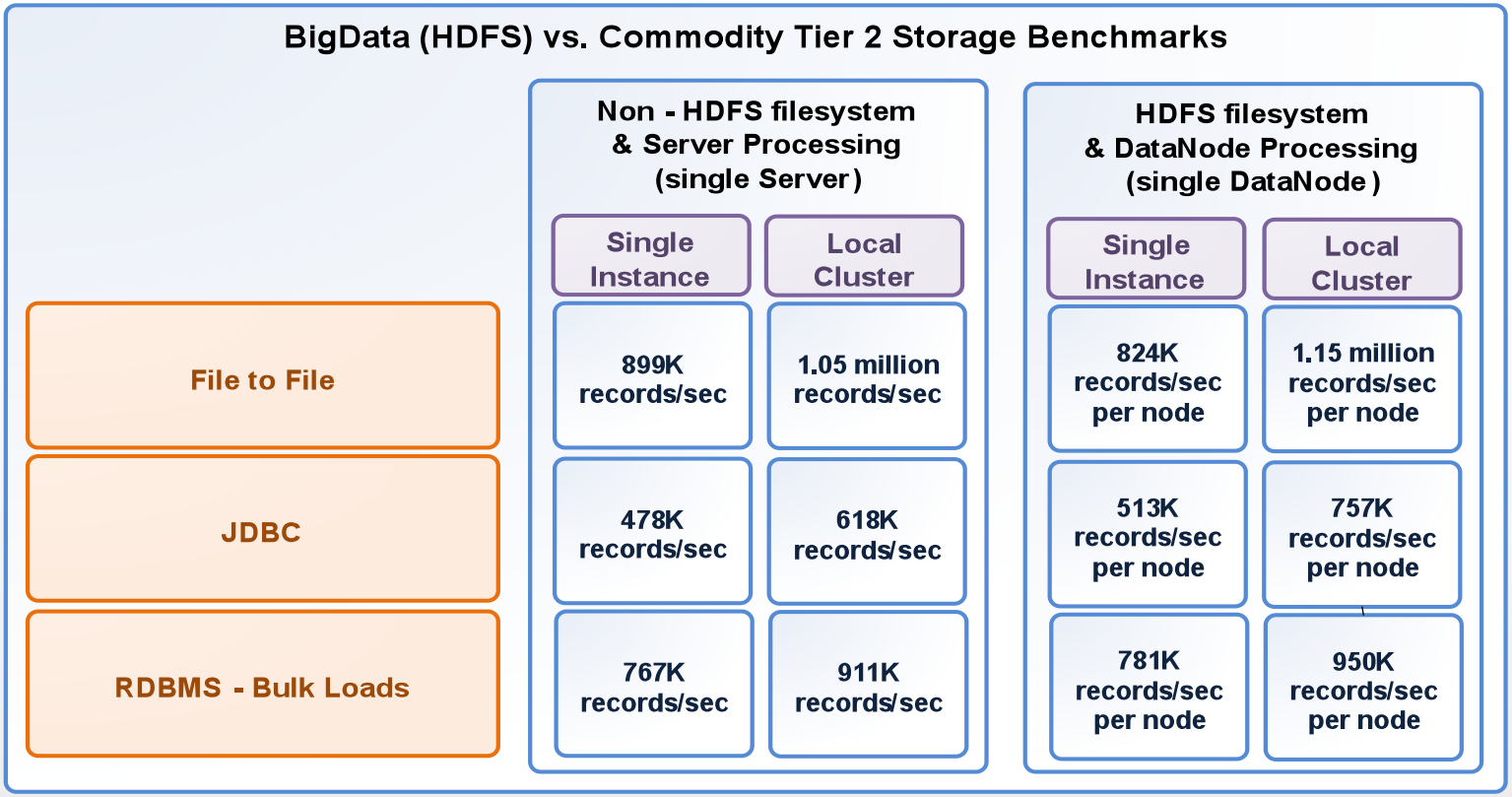
Streaming Data Aggregations – SDA

- 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)

Streaming Data Aggregations – impact on BigData Query Latency



HDSF vs. Tier 2 Storage – Performance Benchmark



StreamHorizon beneficial to BigData filesystem (HDFS)

- Accelerates Hadoop - MapReduce has two main disadvantages (processing is slow & inconvenient to use)
- Hive with StreamHorizon SDA outperforms Impala (Hive based reporting stack usually has higher query latency compared to Impala stack)
- Fine Tune your HDFS via StreamHorizon configuration which enables you to specify size of your HDFS data files.
- Due to aggregation single HDFS file contains even more 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 SDA

- StreamHorizon accelerates ad-hoc queries for Hive and Impala deployments by creating SDA (Streaming Data Aggregates) aggregates on the fly. This significantly reduces I/O of your BigData filesystem (HDFS)
- StreamHorizon reduces memory pressure for Impala (Impala references SDA generated aggregated files)
- StreamHorizon reduces MapReduce processing latency for Hive (processing aggregates takes order of magnitude less time)
- Hive query latency reduced by order of magnitude (function of data volume reduction of your SDA aggregations)

STREAM HORIZON

Q&A