

Understanding Analytical Engines For Networking Analysis Tools

Duncan Pauly

CTO – Edge Intelligence

Agenda

- About the Author
 - Overview
 - Database Dictionary
 - Analytical Engine Descriptions
 - Streaming
 - Row Store
 - Column Store
 - Hadoop
 - Engines Compared
 - Engine Limitations
 - Impact for Network Analytics
-

About the author

- Founder:

- Edge Intelligence: Subscriber Analytics and generic Big Data analytical platform
- CopperEye: High-performance data management
- Zenulta: High-performance event correlation
- Xi Systems: Telecommunication customer care and billing systems

- Patented “tunnel” store database technology

- Holder of numerous other patents related to data management and database design

Overview

- Understanding analytical engines is essential to any analytical product regardless of market
 - Analytical engines greatly influence use case definitions
 - Engine defines use case types
 - at the same time -
 - Limits use case types
 - Network analytics is plagued by the same restrictions – all require an analytics engine
 - Peering analysis
 - DDOS detection
 - Subscriber Analytics
 - Routing analytics
 - This presentation will look at the general available engines and their benefits and detriments. It will also help you think of your current deployments and future analytics purchases to help understand if the product engine is up to the task.
-

Database Dictionary

Term	Definition
Acquisition	Time it takes to add data to the database
Performance	Time it takes to extract data from the database
Query	A desired dataset request
Index	A data structure that helps speed database queries for large datasets
Partitioning	Subdivision of a large table are into multiple smaller parts
Drill down	A subsequent query on a desired dataset

Analytical Engine Options

- Streaming Analytics
- Row Store
- Column Store
- Hadoop



Analytical Engine - Streaming

○ Principle

- Apply queries to incoming data to generate outcomes in real-time
- No retention of historic data

○ Implications

- Immediate query evaluation avoids performance limitations of storage
- Evaluate query results immediately as new data arrives
- Queries must be predefined and use cases known
- Limitations to the query constructs supported
- Prior knowledge of queries effectively creates a point solution
- No ability to retrospectively apply changes to queries
- No ability to drill into the detail behind the analytical outcomes



Analytical Engine – Row Store

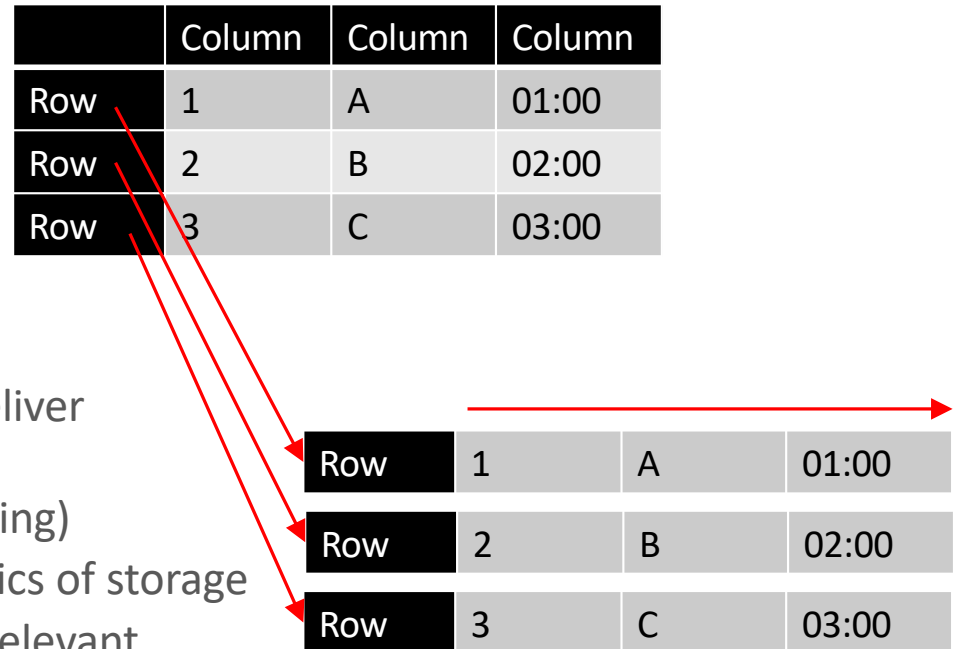
○ Principle

- Store data by row
- Permits fast fetch of rows via indexing

○ Implications

- Good for fetching detailed data
- Relies on partitioning and index design to deliver performance
- Limited data acquisition speed (due to indexing)
- Dependent on the performance characteristics of storage
- Poor for analysis of column subsets (unless relevant indexes exist)
- Indexes create significant storage overhead

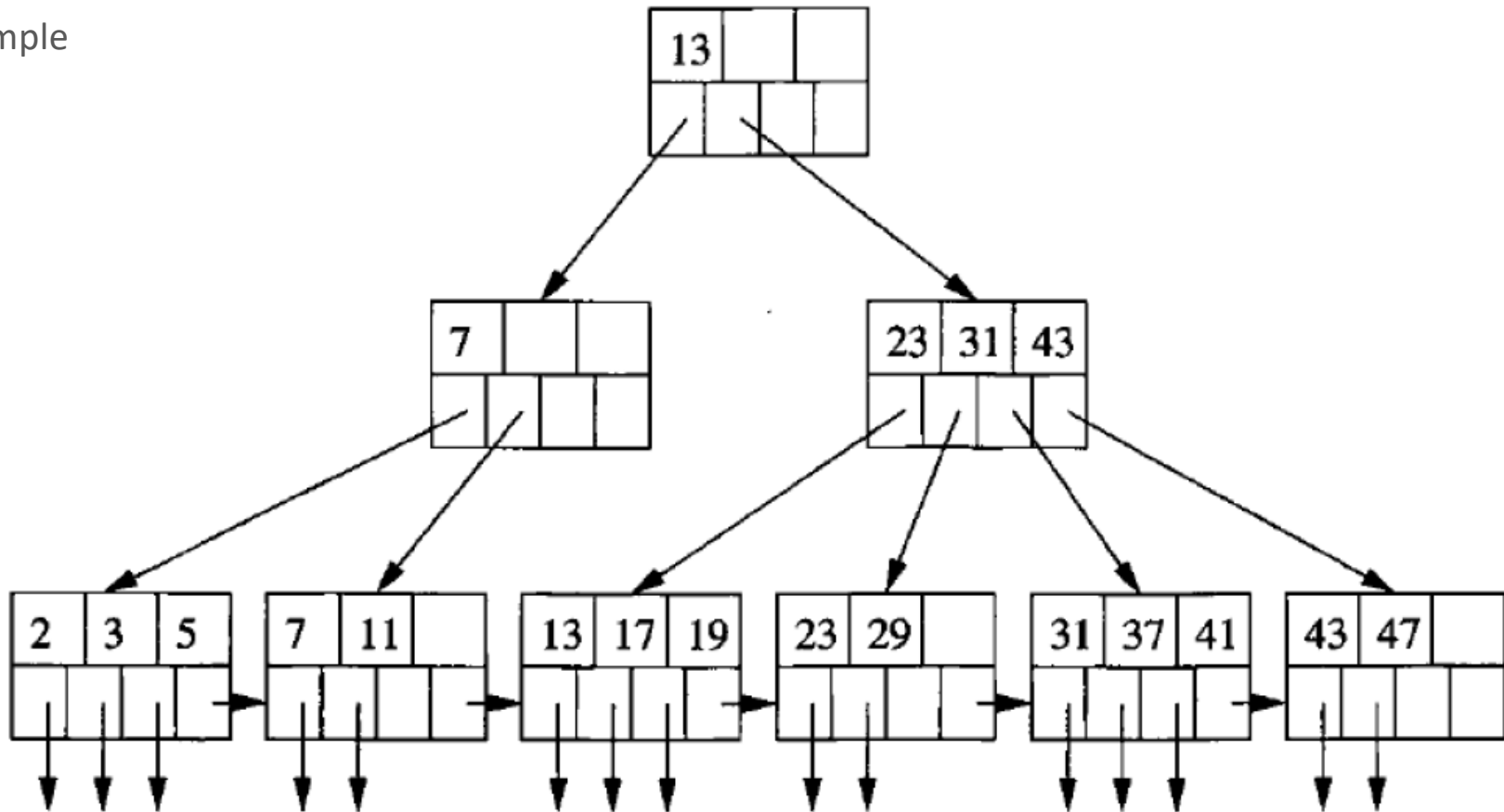
	Column	Column	Column
Row	1	A	01:00
Row	2	B	02:00
Row	3	C	03:00



Row	1	A	01:00
Row	2	B	02:00
Row	3	C	03:00

Indexes

B-Tree Example



Maintaining structures like this slows down data inserts and updates

Analytical Engine – Column Store

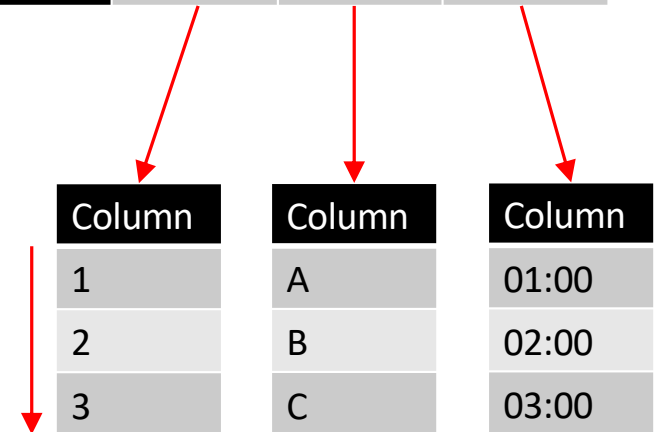
○ Principle

- Store data by column
- Permits fast scan of column subsets

○ Implications

- Good for analysis of column subsets for aggregate use cases
- Good data acquisition speed
- Good data compression rates
- Dependent on the performance characteristics of storage
- Poor for drill downs into the detail behind the analytical result sets

	Column	Column	Column
Row	1	A	01:00
Row	2	B	02:00
Row	3	C	03:00



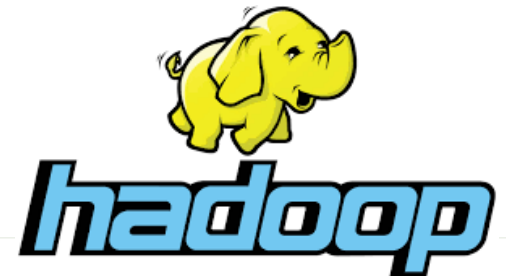
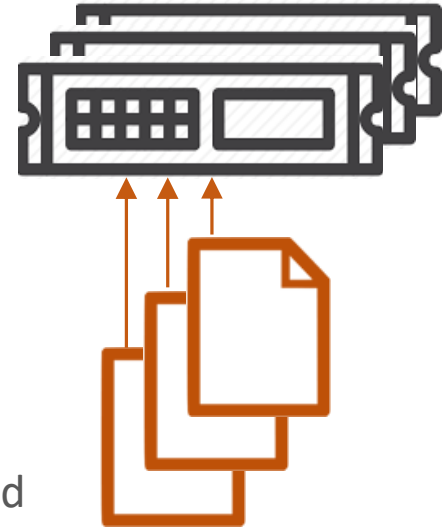
Analytical Engine – Hadoop

- Principle

- Parallel file scanning and partitioning

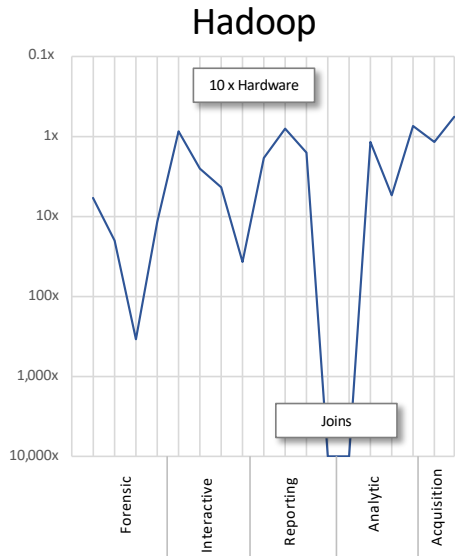
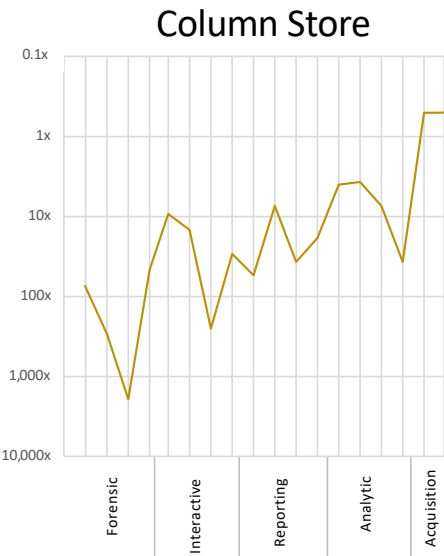
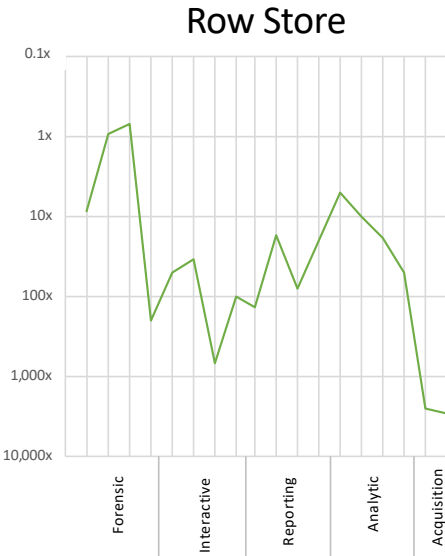
- Implications

- Relies on large hardware estates to deliver performance
- Offers good data acquisition speed
- Performance depends on the hardware resources available and concurrent workloads (because of IO and network saturation)
- Some design decisions are required which will favour certain use cases over others (eg. file type and partitioning scheme)



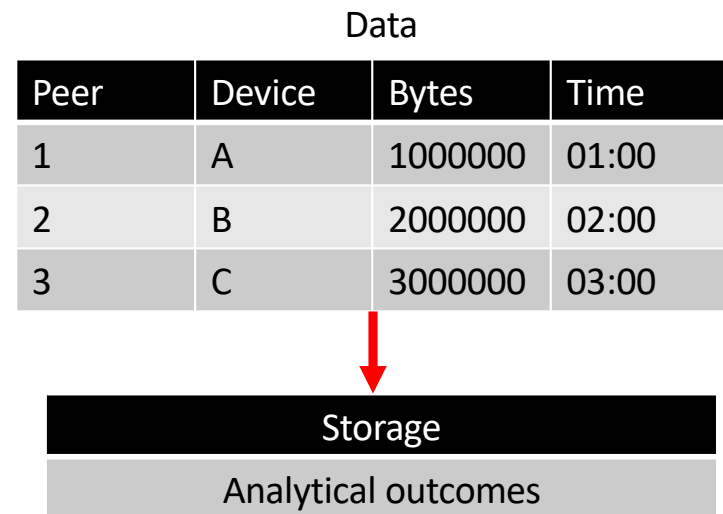
Analytical Engine – Performance profiles compared

Queries and Data Acquisition Performance



Streaming Engine – Impact on Network Analytics

- Streaming engine is good for Network analytics requiring quick acquisition and query. Limitation is the inflexibility of queries if further analysis is required to understand a root cause.
 - **Good for:**
 - DDOS detection – due to real time telemetry
 - Bad actor detection
 - Real-time dashboards
 - **Bad for:**
 - Ad-hoc queries
 - Any unanticipated use cases
 - Unable to drill into detail behind analytical outcomes



Row Store – Impact on Network Analytics

- Row store engine is good for needle-in-a-haystack forensic analysis provided the correct indexes are implemented. Aggregates generally perform poorly.

- **Good for:**

- Subscriber specific analysis
- Narrow time-window analysis

- **Quick access to all data in a row**

- Single row read for specific peer and timestamp

- **Slow access for 'column' data for aggregate**

- Depends on indexes available

- **Slow acquisition (adding to database)**

- Acquisition rate can be slow
 - because of index maintenance
- Indexes require significant storage overhead
- Performance relies on storage characteristics and database design in particular indexing & partitioning

Data

Peer	Device	Bytes	Time
1	A	1000000	01:00
2	B	2000000	02:00
3	C	3000000	03:00




Row Store
1,A,1000000,01:00
2,B,2000000,02:00
3,C,3000000,03:00

Column Store - Impact on Network Analytics

- Column store engine is good for aggregate use cases but drilling down into the detail can be slow because of the need to locate and reconstruct rows.
 - **Good for aggregate use cases:**
 - Peering analysis – e.g. traffic by peer XYZ
 - Network load distribution – eg. by Device
 - Time based volume analysis and trending
 - **Acquisition and compression**
 - Good data acquisition speed and compression rates
 - **Poor performance for drill downs behind aggregate results sets**
 - Requires row reconstruction
 - **Performance dependent on storage characteristics**
 - NAS, SAN, SSD, HDD

Data

Peer	Device	Bytes	Time
1	A	1000000	01:00
2	B	2000000	02:00
3	C	3000000	03:00



Column Store

1,2,3
A,B,C
1000000,2000000,3000000
01:00,02:00,03:00

Hadoop – Impact on Network Analytics

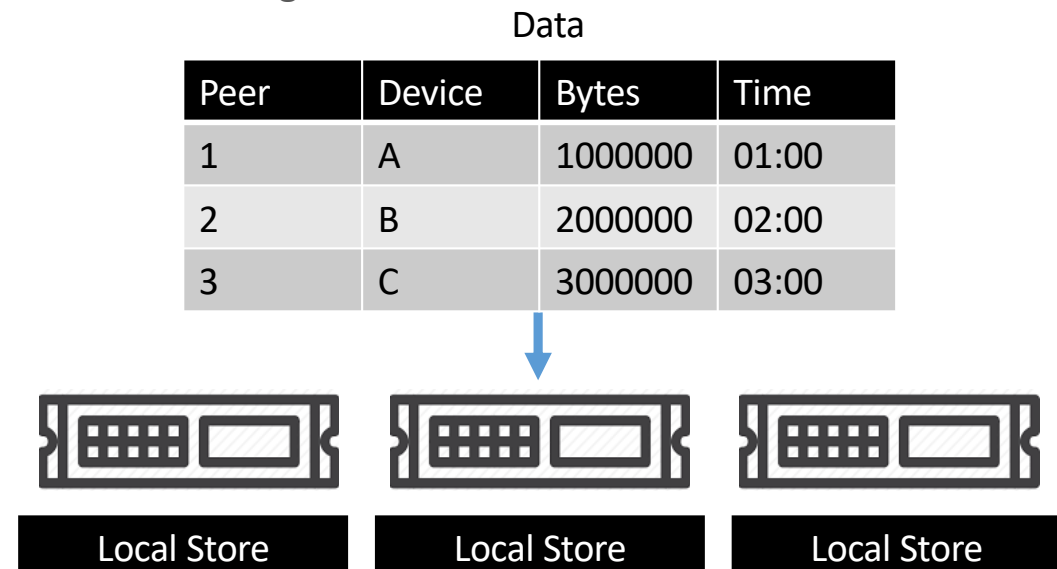
- Hadoop engine can support many non-real-time use cases - but at a high hardware cost. Typically the hardware footprint is an order of magnitude greater compared to best in class above.

- **Good for:**

- Data acquisition
- Performance largely depends on available hardware

- **Bad for:**

- Certain use cases will be favoured over others due to design decisions
- Real-time use cases
- Performance heavily impacted by concurrent workloads



Analytical Engines – Summary

- Network telemetry volume grows in line with network volume growth:
 - Netflow/IPFix, IP Lease, DNS, BGP, IGP, etc.
 - Can impact Hadoop and Column stores adversely
 - Each analytical engine excels in certain tasks and struggles with other tasks
 - Essentially, each becomes a point solution for a subset of use cases:
 - Streaming requires prior knowledge of queries
 - Both Row and Column stores have asymmetric performance profiles & limitations
 - Hadoop requires significantly larger hardware investment to deliver equivalent performance
 - Understanding the analytical engine of any analytical tool is crucial to understand if it can deliver the desired results
-

Thank You!

duncan.pauly@edgeintelligence.com