COMPUTERWORLD SNIA

SNW

Driving Innovation Through the Information Infrastructure

SPRING 2011

Big Data

How New Data Analytics Systems will Impact Storage

John Webster

Senior Partner

Evaluator Group

There is no question about whether or not traditional data warehousing will be transformed into "Big Data" analysis applications. Some vendors are already responding to the burgeoning need for IT processes that converge traditional database data with data from multiple disparate sources that include online, wireless, and sensory resources-hence the term "Big Data." This need is being seen across industry segments, but is currently most prominent in healthcare, government, and retail. What storage professionals need to understand now is how the Big Data wave will impact data storage

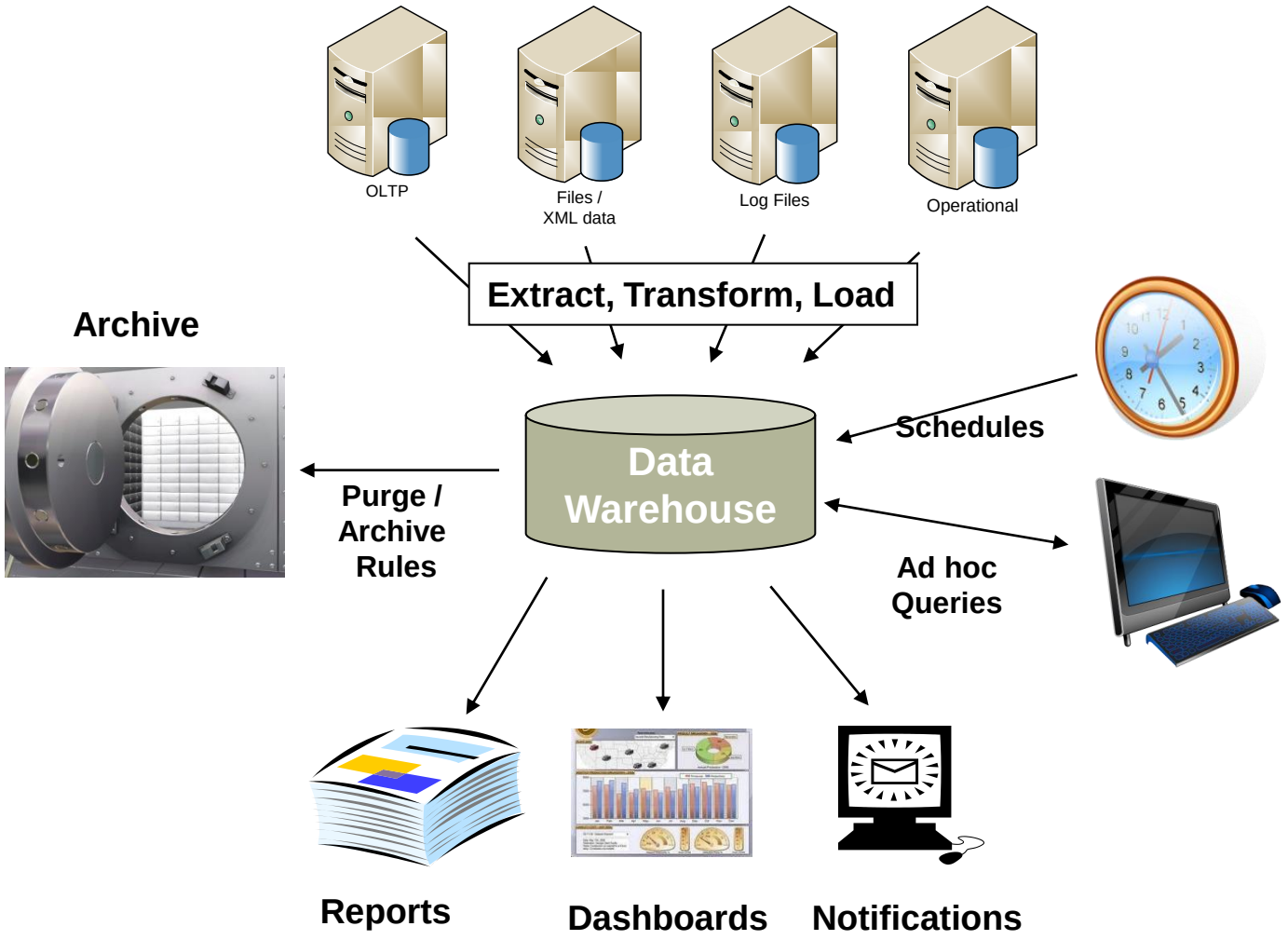
Outline

- Define Big Data Analytics
- New Analytics Engines
- The New Analytics Engines and Storage
- Is Big Data in Your Future?

Which Big Data am I Talking About?

- Big Data Storage
  - Systems that store really big (as in humungous) amounts of data
- Big Data Analytics
  - Systems that use new analytics processes to crunch really big amounts of data from multiple sources and deliver information in real or near real time
- Big Data Storage that supports Big Data Analytics

Traditional Data Warehouse Process Flow



Big Data Analytics Fundamentals

- Multiple, high volume data sources
- Convergence yields new types of information, often leveraged for competitive advantage
- Time to information is often critical, therefore

LATENCY IS THE ENEMY

Big Data Analytics Platforms

- In-Memory Database (IMDB)
- Hadoop/MapReduce and Derivatives
- StreamSQL
- Complex Event Processing (CEP)
- SciDB
- Operational Analytics
- …and a growing list of specialized tools
  - Trendril TREE (example)

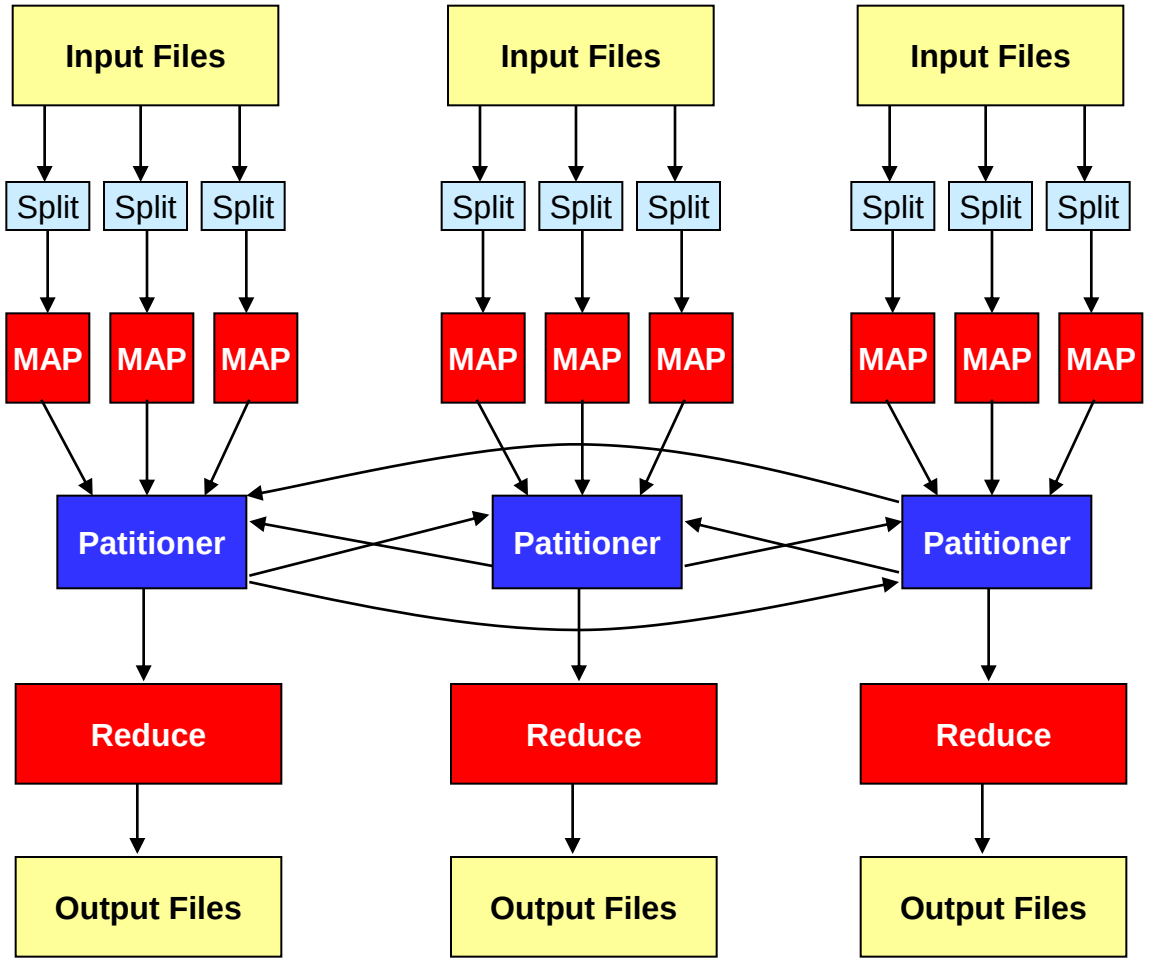
In-Memory Database

- Data stored in main memory
- Reduce latency by minimizing I/Os to disk
- But, increased exposure to data loss if system fails
- Therefore, data persistence can be added in the form of NVRAM, SSD, and hybrid memory/disk architectures
- Cost is a limiting factor

MapReduce Essentials

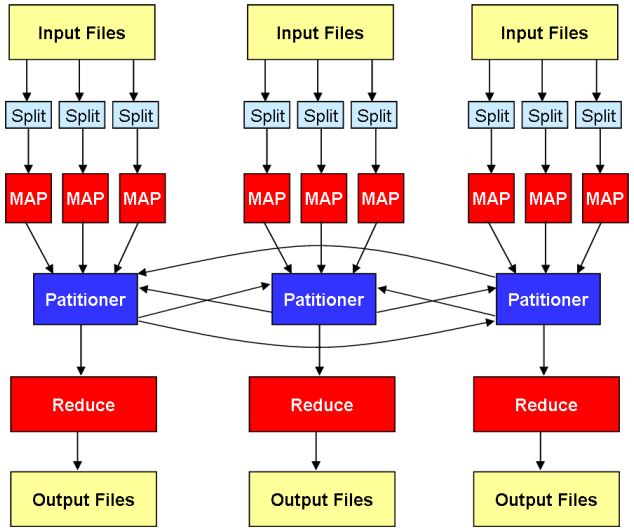
- Programming model originally described by Google in 2004
- Extract information quickly from very large, disparate data sets (aka Big Data).
- Takes data input and divides it into smaller segments to be sent to clustered processing nodes (the “map” step)
- Nodes process assigned segments in parallel and return results (the “reduce” step)
- Hadoop MapReduce: Apache open source project

MapReduce Essentials

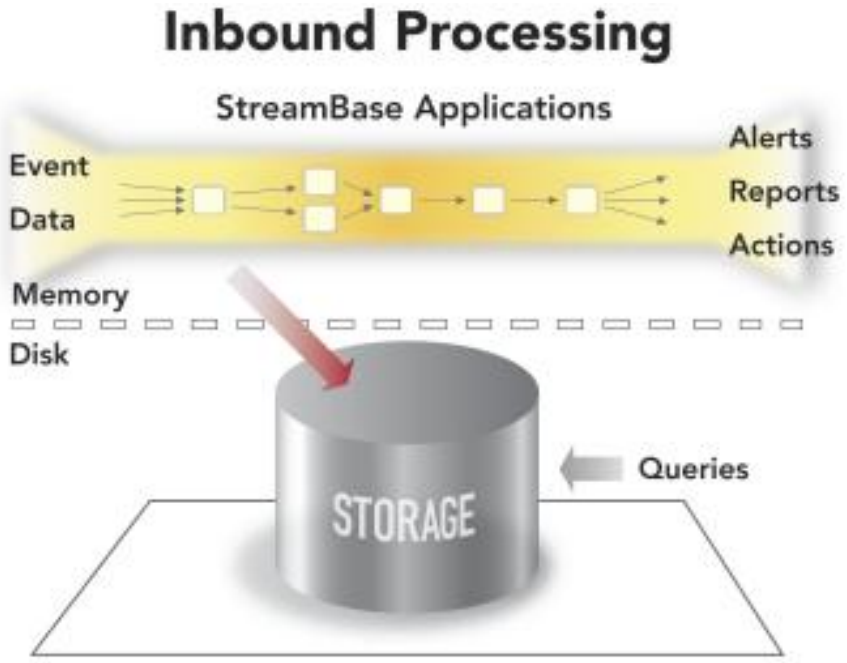


MapReduce Characteristics

- Highly scalable
- Multiple data sources
- Parallelized inputs and processing flow reduces latency
- Commodity components reduce cost
- Shared Nothing (including storage)

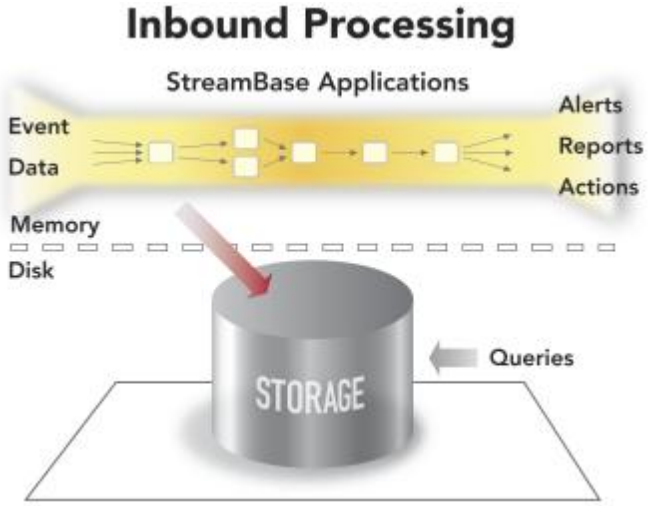


StreamSQL and Complex Event Processing (CEP)

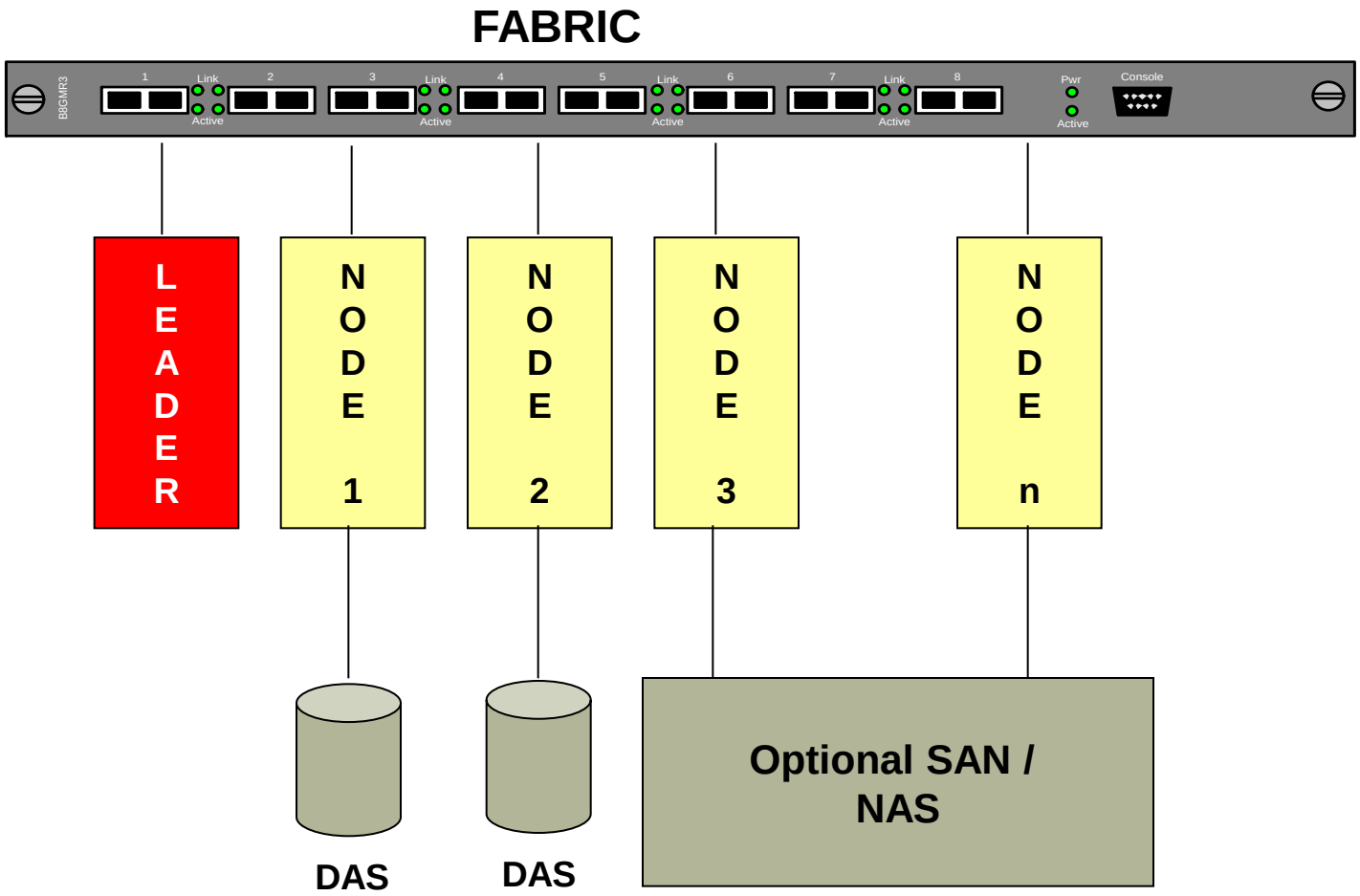


StreamSQL Characteristics

- Multiple concurrent data sources
- Very low latency
- Commodity components
- Storage is a repository

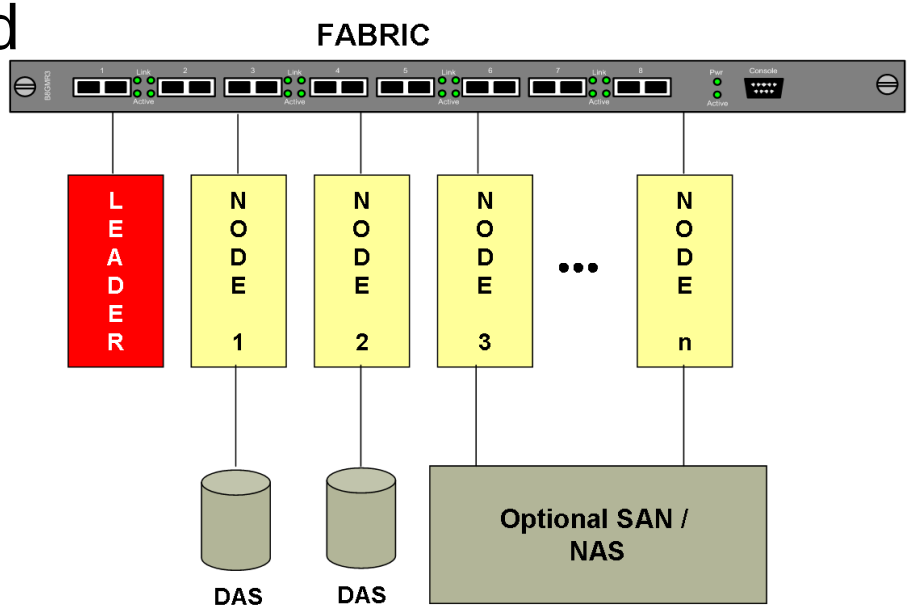


Shared Storage for Shared Nothing

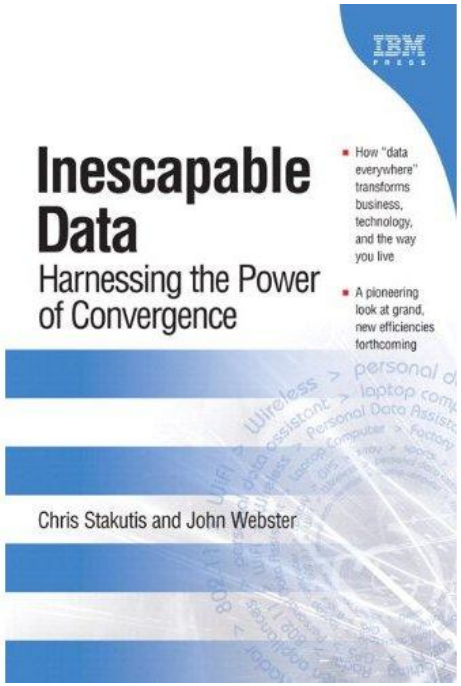


Shared Storage for Shared Nothing

- Node-based data mirrored to back-end SAN or NAS
- Reduces latency for queries that span nodes
- Enhances system availability and data protection



2004: The Power of Convergence



- Virtually anything can be a “data emitter”
- The convergence of mutliple data source can yield new types of information
- Time to information will often be critical

Stride Rite CEO Vision 2004

- Wireless device for measuring the foot
- Data used to select proper shoe size or (future) produce the shoe on-site from parts
- **Data shared with supply chain vendors, non-competing retailers, healthcare providers and medical researchers**

Realizing the Vision 2011 - Podiatric Data Collection

(Big-foot Data?)

Where Are We Going?

- The Internet of Things
- Private and Public Data Clouds
- Omniscient Retailers
- Brain Implants
- …and much more

(Is the Singularity really near?)

Summary

- Big Data Analytics is about the convergence of multiple data streams in real or near-real time yielding new information sources.
- Storage, and particularly shared storage, in the context of big data analytics is often seen as an avoidable source of latency.
- Big Data Analytics is the next storage frontier.