NextGen Infrastructure for Big Data

Anil Vasudeva
President & Chief Analyst
imex@imexresearch.com
408-268-0800

IMEX RESEARCH.COM

Abstract

IMEX RESEARCH.COM

- **NextGen Infrastructure for Big Data**
  - This session will appeal to Business Planning, Marketing, Technology System Integrators and Data Center Managers seeking to understand the drivers behind the demand for and rise of Big Data.
  - **Abstract**
  - The internet has spawned an explosion in data growth in the form of data sets, called Big Data, that are so large they are difficult to store, manage and analyze using traditional RDBMS which are tuned for Online Transaction Processing (OLTP) only. Not only is this new data heavily unstructured, voluminous and streams rapidly and difficult to harness but even more importantly, the infrastructure cost of HW and SW required to crunch it using traditional RDBMS, to derive any analytics or business intelligence online (OLAP) from it, is prohibitive.
  - To capitalize on the Big Data trend, a new breed of Big Data technologies (such as Hadoop and others) many companies have emerged which are leveraging new parallelized processing, commodity hardware, open source software and tools to capture and analyze these new data sets and provide a price/performance that is 10 times better than existing Database/Data Warehousing/Business Intelligence Systems.
  - **Learning Objectives**
  - The presentation will illustrate the existing operational challenges businesses face today using RDBMS systems despite using fast access in-memory and solid state storage technologies. It details how IT is harnessing the emergent Big Data to manage massive amounts of data and new techniques such as parallelization and virtualization to solve complex problems in order to empower businesses with knowledgeable decision-making.
  - It lays out the rapidly evolving big data technology ecosystem - different big data technologies from Hadoop, Distributed File Systems, emerging NoSQL derivatives for implementation in private and hybrid cloud-based environments, Storage Infrastructure Requirements to Store, Access, Secure, Prepare for analytics and visualization of data while manipulating it rapidly to derive business intelligence online, to run businesses smartly.

2

Big Data in IT Industry Roadmap

IMEX RESEARCH.COM

IT Industry Roadmap

Analytics – BI

Predictive Analytics - Unstructured Data

From Dashboards Visualization to Prediction Engines using Big Data.

Cloudization

On-Premises > Private Clouds > Public Clouds

DC to Cloud-Aware Infrast. & Apps. Cascade migration to SPs/Public Clouds.

Automation

Automatically Maintains Application SLAs

(Self-Configuration, Self-Healing©IMEX, Self-Acctg. Charges etc.)

Virtualization

Pools Resources. Provisions, Optimizes, Monitors

Shuffles Resources to optimize Delivery of various Business Services

Integration/Consolidation

Integrate Physical Infrast./Blades to meet CAPSIMS ©IMEX

Cost, Availability, Performance, Scalability, Inter-operability, Manageability & Security

Standardization

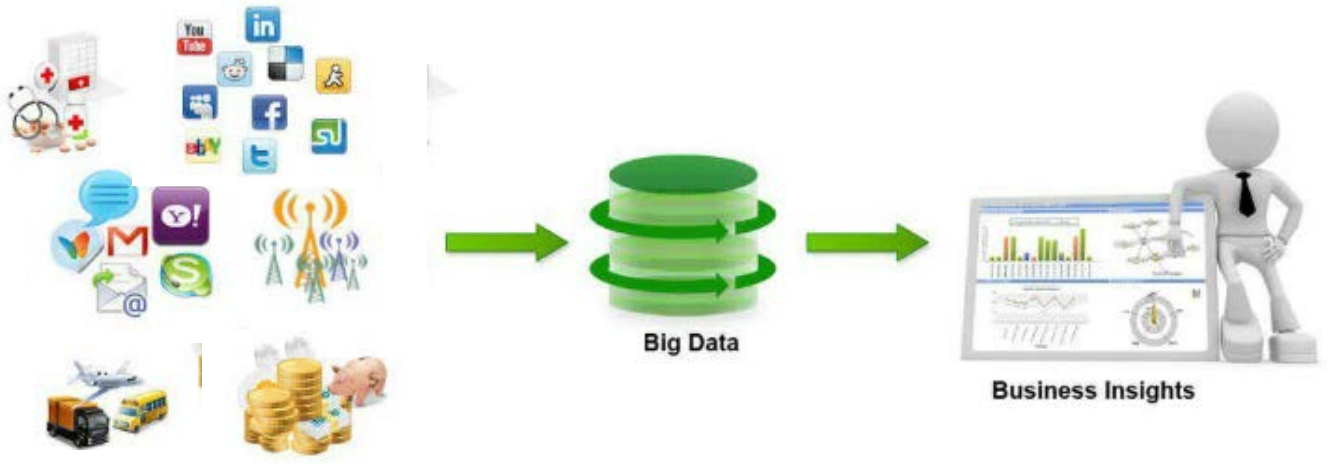
Standard IT Infrastructure- Volume Economics HW/Syst SW

(Servers, Storage, Networking Devices, System Software (OS, MW & Data Mgmt. SW)

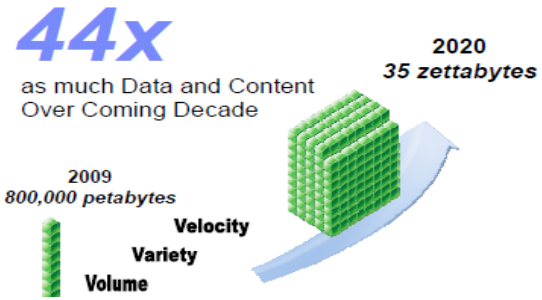
Harnessing Big Data for Business Insights

IMEX RESEARCH.COM

Data Sources → **Big Data Infrastructure** → **Business Insights**

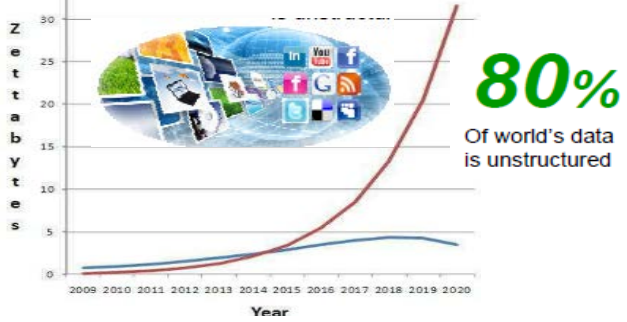


Information is at the center of New Wave of opportunity



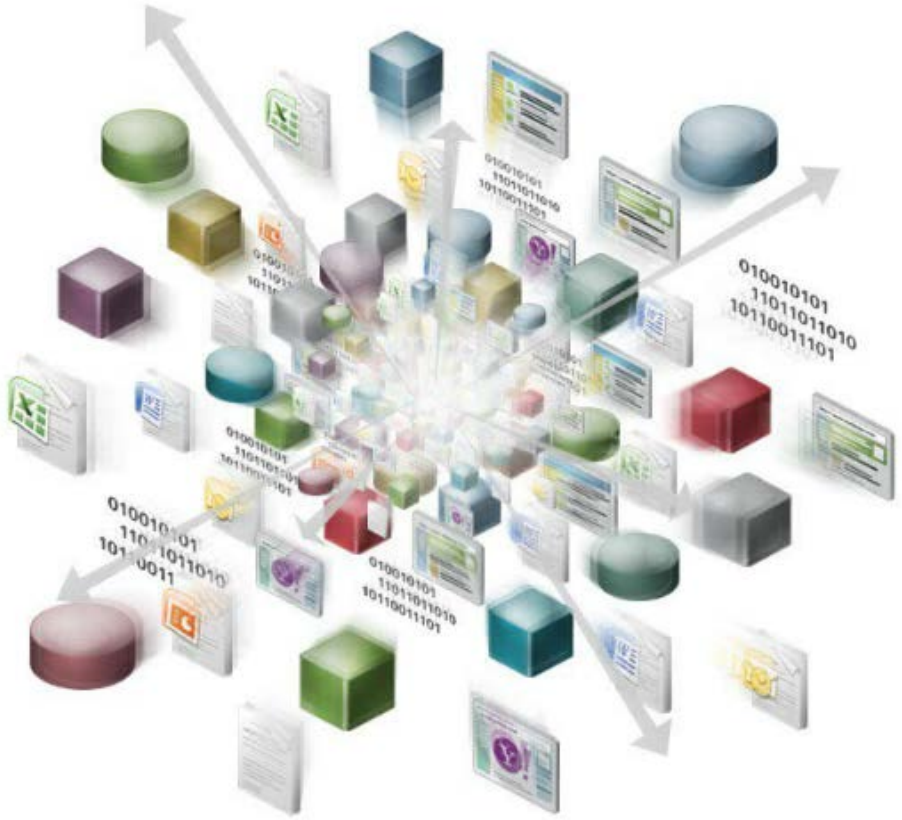
Majority of data growth is being driven by unstructured data and billions of large objects

80% of world's data is unstructured driven by rise in Mobility devices, collaboration machine generated data.



Corporate Need: Business Perf... Optimization

IMEX RESEARCH.COM

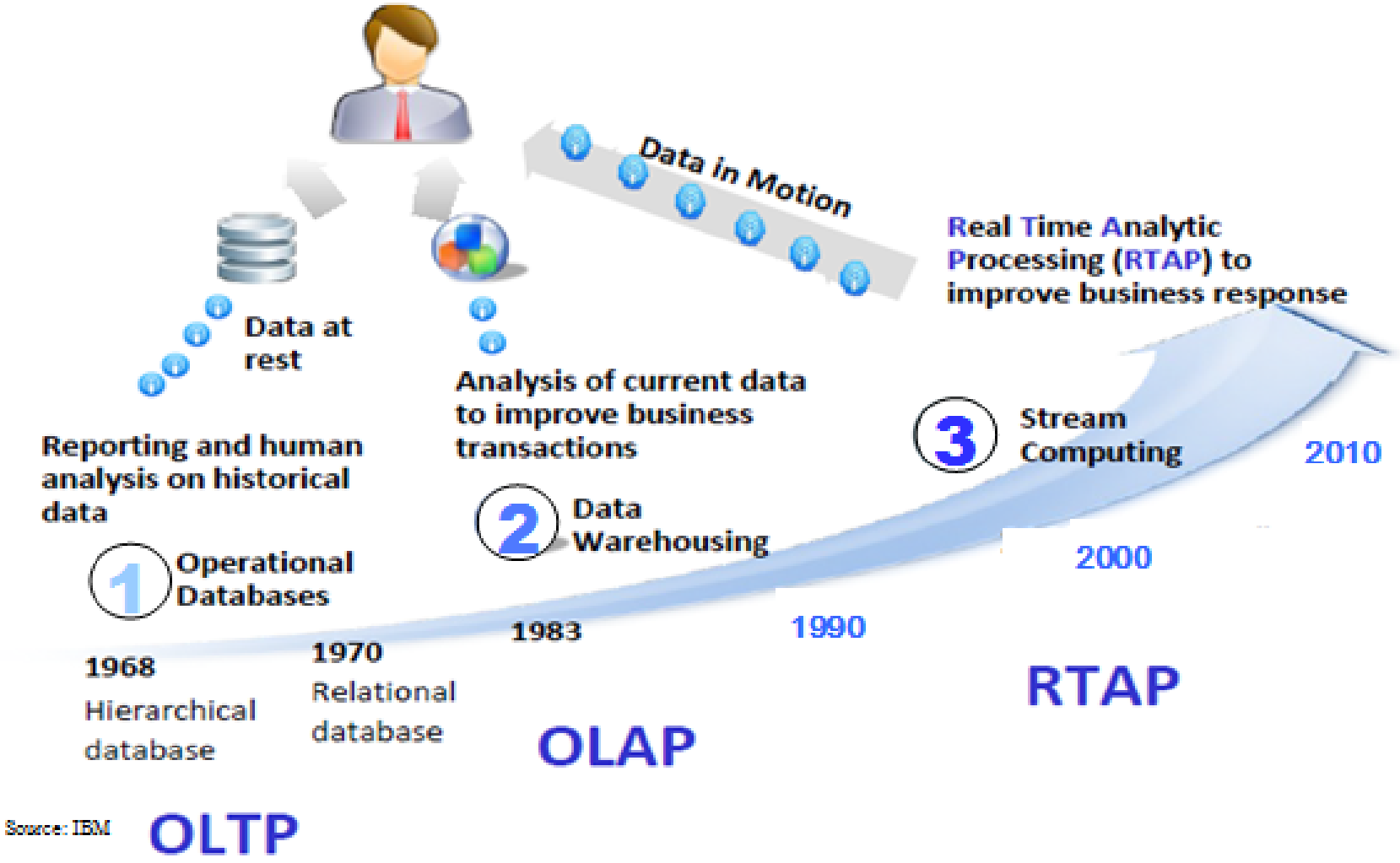


Unstructured Big Data can provide Next Gen Analytics to help businesses make informed, better decision in:

- Product Strategy
- Targeting Sales
- Just-In-Time Supply-Chain Economics
- Business Performance Optimization
- Predictive Analytics & Recommendations
- Country Resources Management

Corporate Need: Real Time Analytics

IMEX RESEARCH.COM



Source: IBM

Corporate Need: Business Insights

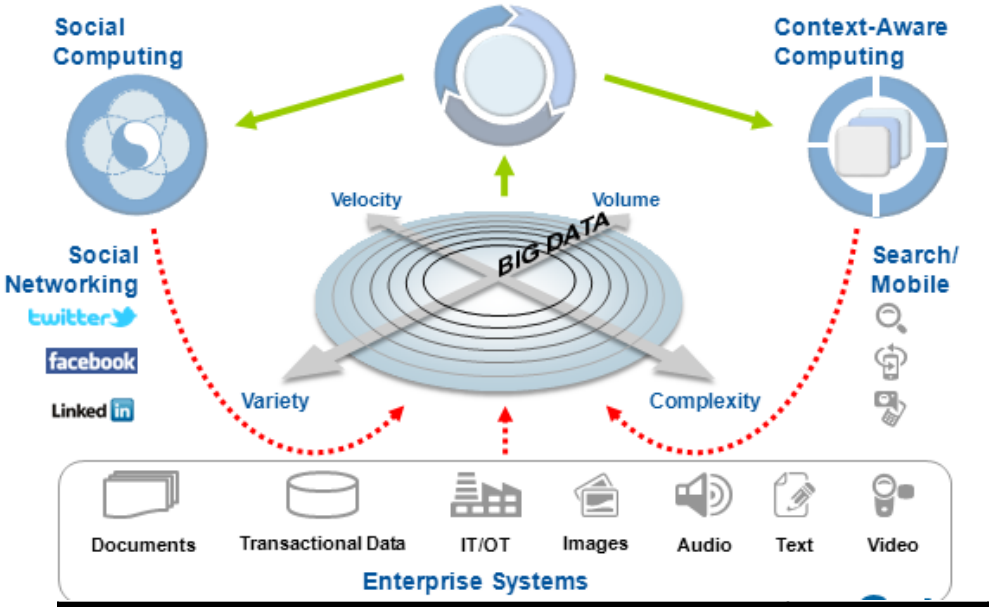
IMEX RESEARCH.COM

Item	Issue	Solution
Store	**Information Exploding** Volume: Digital Content doubling every 18 months. Velocity: >80% growth driven from unstructured data. Variety: sources of data changing	**A unified information/content storage methodology** that enables users to manage the volume, velocity and variety of information from multiple sources
Manage	**Complexity in "managing" information.** - Need to classify, synchronize, aggregate, integrate, share, transform, profile, move, cleanse, protect, retire	**A solution portfolio of tools and services** to manage all types of information in a hybrid storage environment
Analyze	**Current solutions limited to BI** tools focused on structured and lagging information	Build/buy packaged **Real-Time Predictive Analytical Solutions** for unstructured analytics tools
Collaborate	**Multiple access methods** needed to meet needs of a diverse audience.	**Centralized share, collaborate** and act on insights anytime, anywhere on any device.
Model/ Adapt	Ability to **understand how the information impacts the business**. How to transfer to action.	**Model Information on current operations w/potential strategy impact.** Leverage Tech. to adapt.

Opportunity: Converting Big Data Deluge into Predictive Analytics & Insights

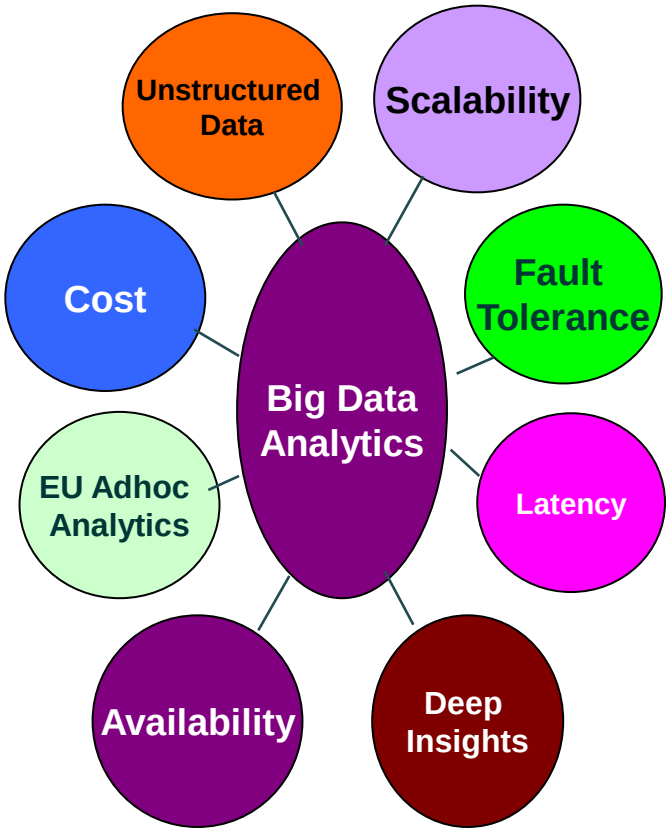
IMEX RESEARCH.COM

Big Data Predictive Analytics



Personal Location Services Data Generated by	*TB/Year*
Navigation Devices	600
Navigation Apps on Phones	20
Smart Phone Opt-In Tracking	1000
Geo Targeted Ads	20
People Locator (Emergency Calls/Search..)	10
Location based Services (e.g.Games)	5
Other	45
Total (Est.)	1700

Issues with Existing RDBMS



Key Issues with RDBMS Technologies

Handling Mixed Unstructured Data

- RDBMs don't handle non-tabular data
  (Notorious for doing a poor job on recursive data structure)

Legacy Archaic Architecture

- RDBMS don't parallelize well to accommodate commodity HW clusters

Speed

- Seek time of physical Storage has not kept pace with network speed improvements

Scale

- Difficult to scale-out RDBMS efficiently – Clustering beyond few servers notoriously hard

Integration

- Data processing tasks need to combine data from non-related sources, over a network

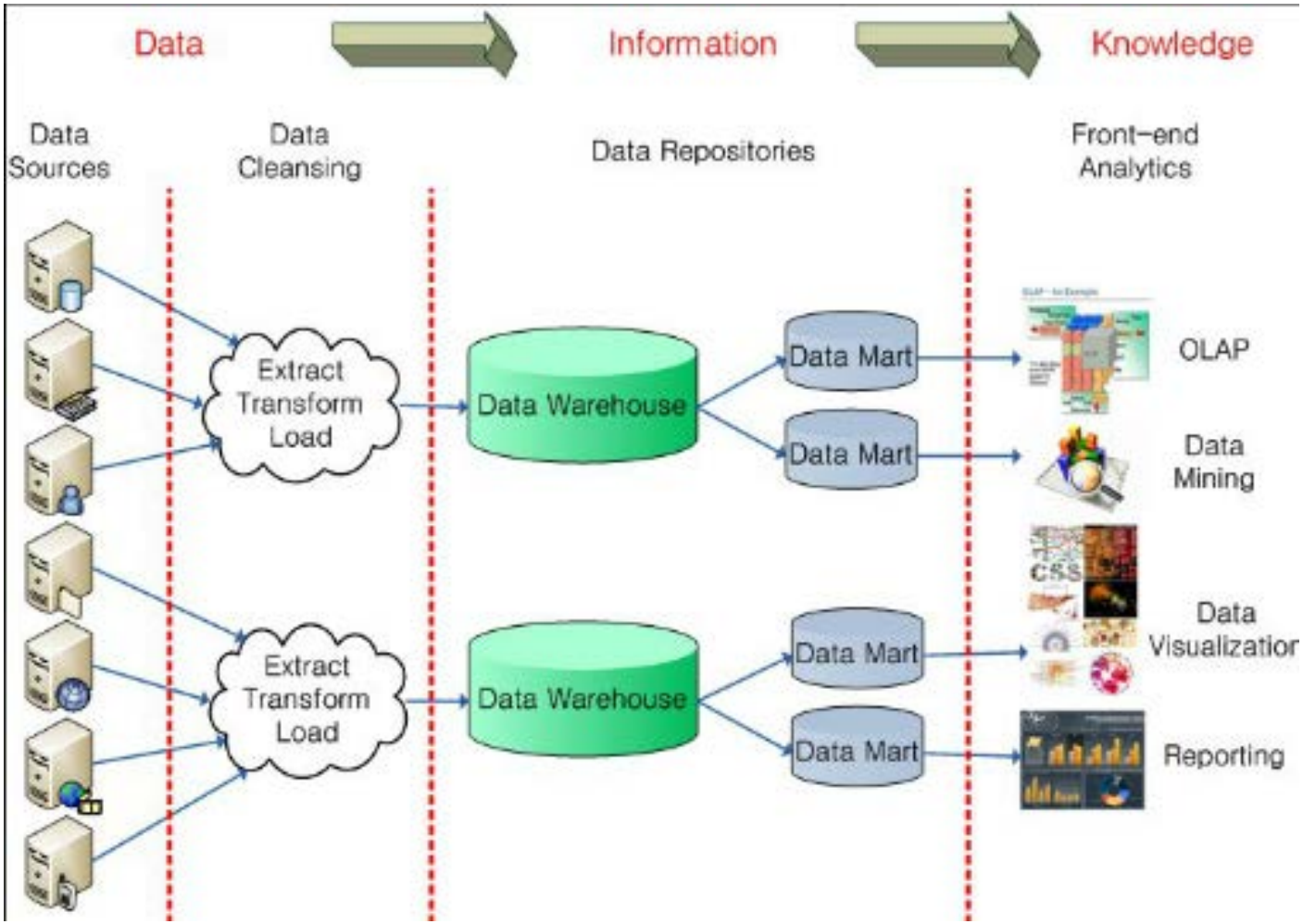
Volume

- Data volumes have grown from 10s GB >100s TB > PBs in recent years. Existing Tabular RDBMS can't handle such large DBs

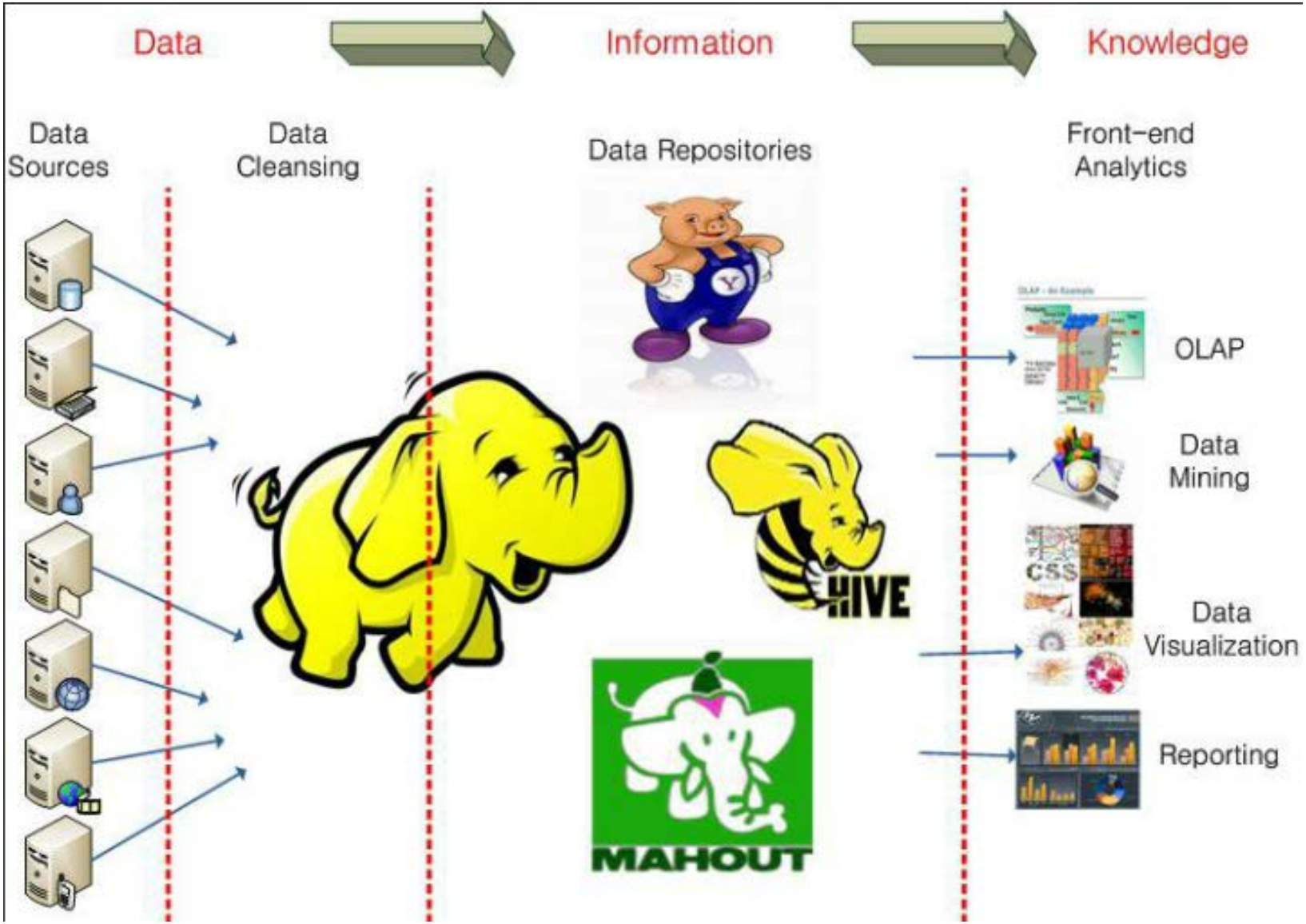
Issues with Existing RDBMS

IMEX RESEARCH.COM

Present RDBMS struggling to Store & Analyze Big Data



Big Data - Database Solutions



Big Data – The New Face of DBs

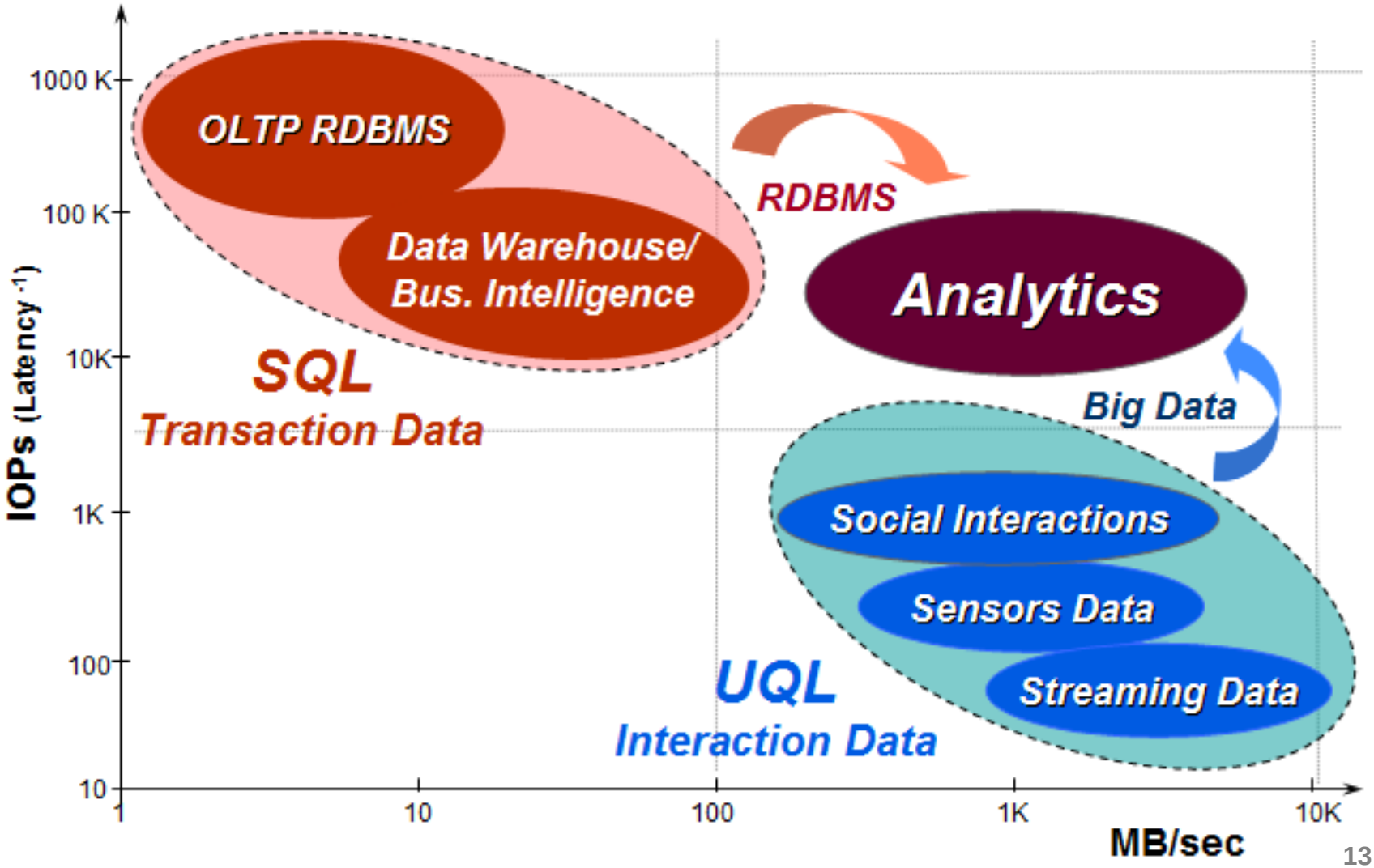
IMEX RESEARCH.COM

Big Data Paradigm - The New face of DB Systems

- Adopts Schema-Free Architecture
- Can do away with Legacy Relational DB Systems
  - Some data have sparse attributes, do not need relational property
- Key Oriented Queries
  - Some data stored/retrieved mainly by primary key, w/o complex joins
- Trade-off of Consistency, Availability & Partition Tolerance
- Scale Out, not up, - Online Load balancing cluster growth

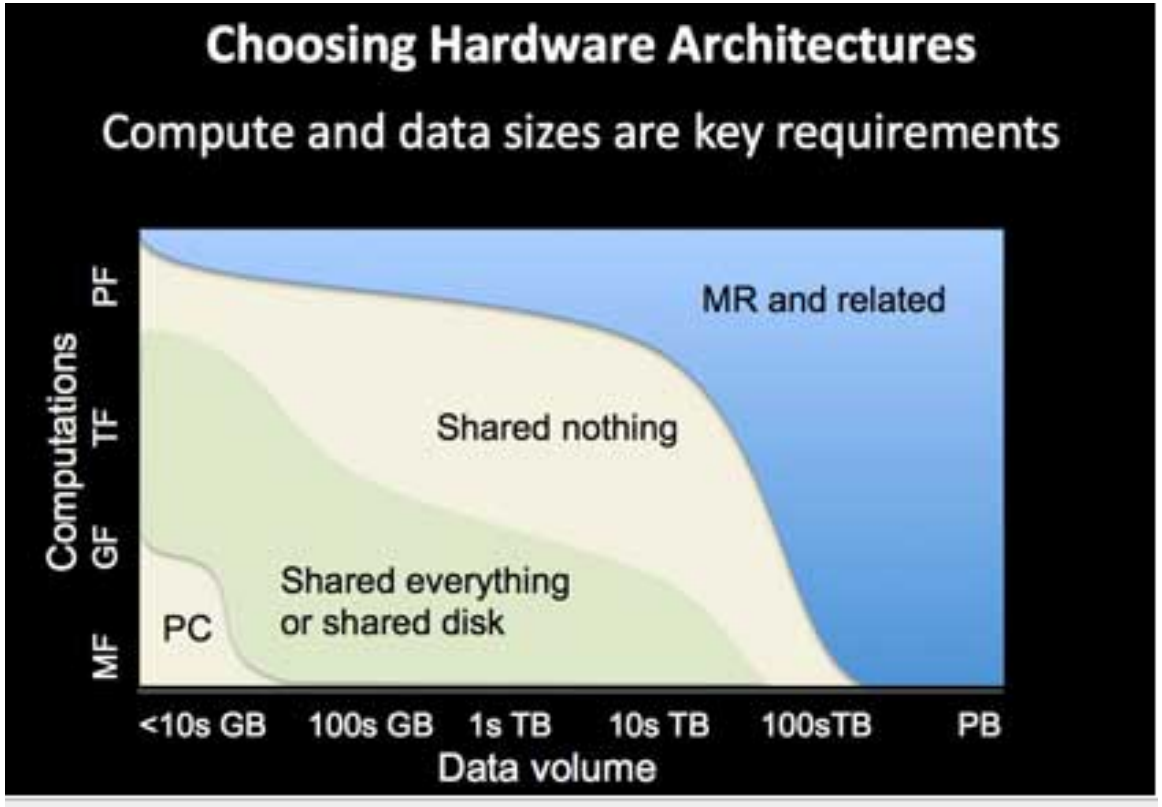
Analytics – The Next Frontier in IT

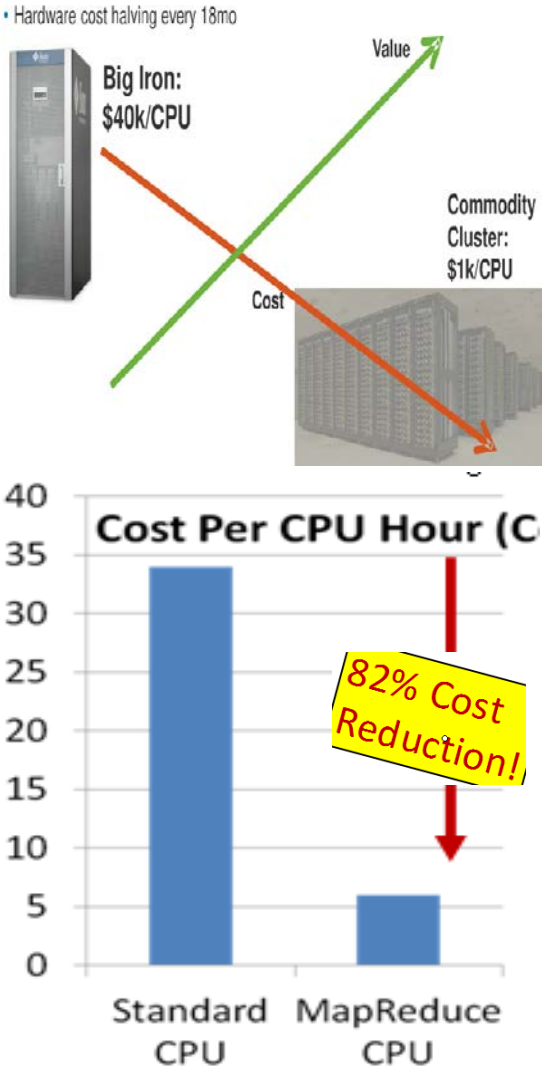
IMEX RESEARCH.COM



Key Innovations: HW Technologies

IMEX RESEARCH.COM

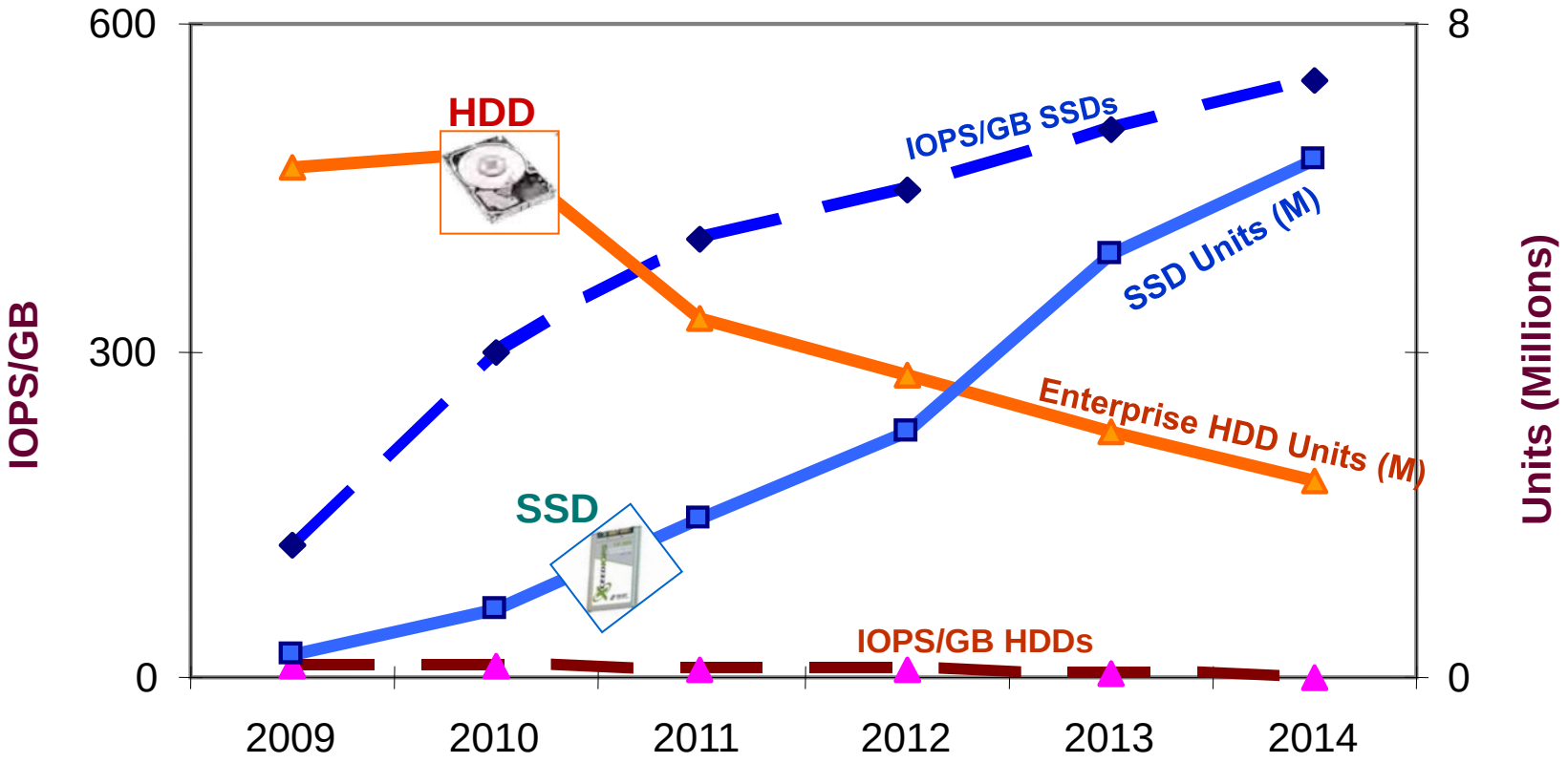




Key Innovations – Solid State Storage

IMEX RESEARCH.COM

Storage - IOPS/GB & Price Erosion - HDD vs. SSDs



Note: 2U storage rack, • 2.5" HDD max cap = 400GB / 24 HDDs, de-stroked to 20%, • 2.5" SSD max cap = 800GB / 36 SSDs

Key to Database performance are random IOPS. SSDs outshine HDD in IO price/performance – a major reason, besides better space and power, for their explosive growth.

Source: IMEX Research SSD Industry Report ©2011

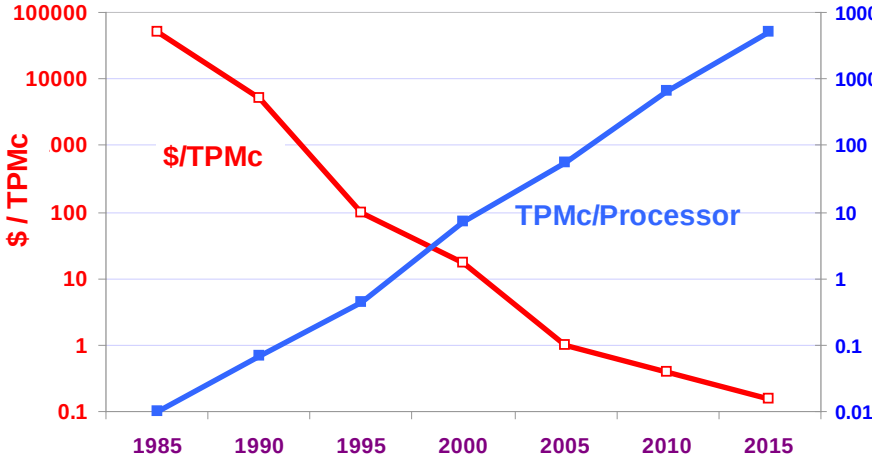
15

Innovations – DB SW Technologies

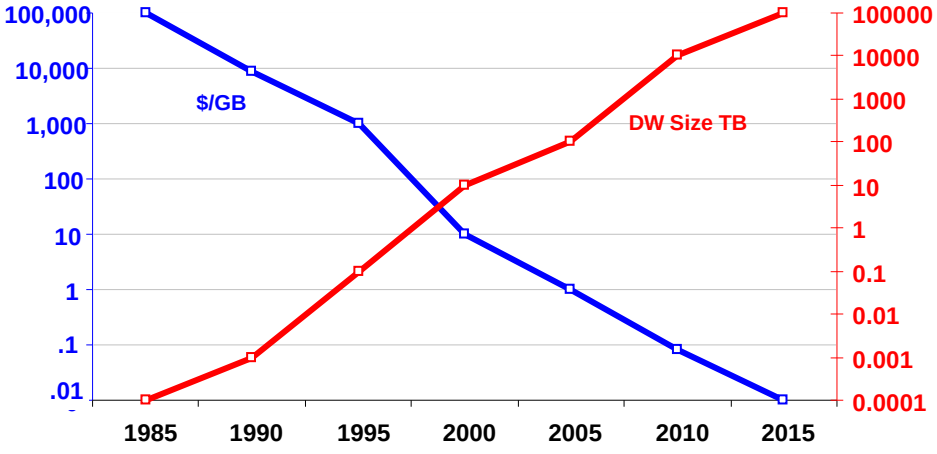
IMEX RESEARCH.COM

Tech Innovation	1985	1990	1995	2000	2005	2010	2015
OLTP Transactions DB SW	Rows Locking	Optimizer	Parallel Query	Clustering	XML	Grid	Open Source / Hadoop
OLAP- Analytics DB SW	Indexing	Partitioning	Columnar	Materialized View	Bit Mapped Index	In-Memory	Query Binding
Hardware	32 bit	SMP	NUMA	64 bit	Multi-core/ Blades	Flash	MPP
Big Data					Multi-core	Columnar In-Memory	MPP Visualization

OLTP Database Innovation Progress

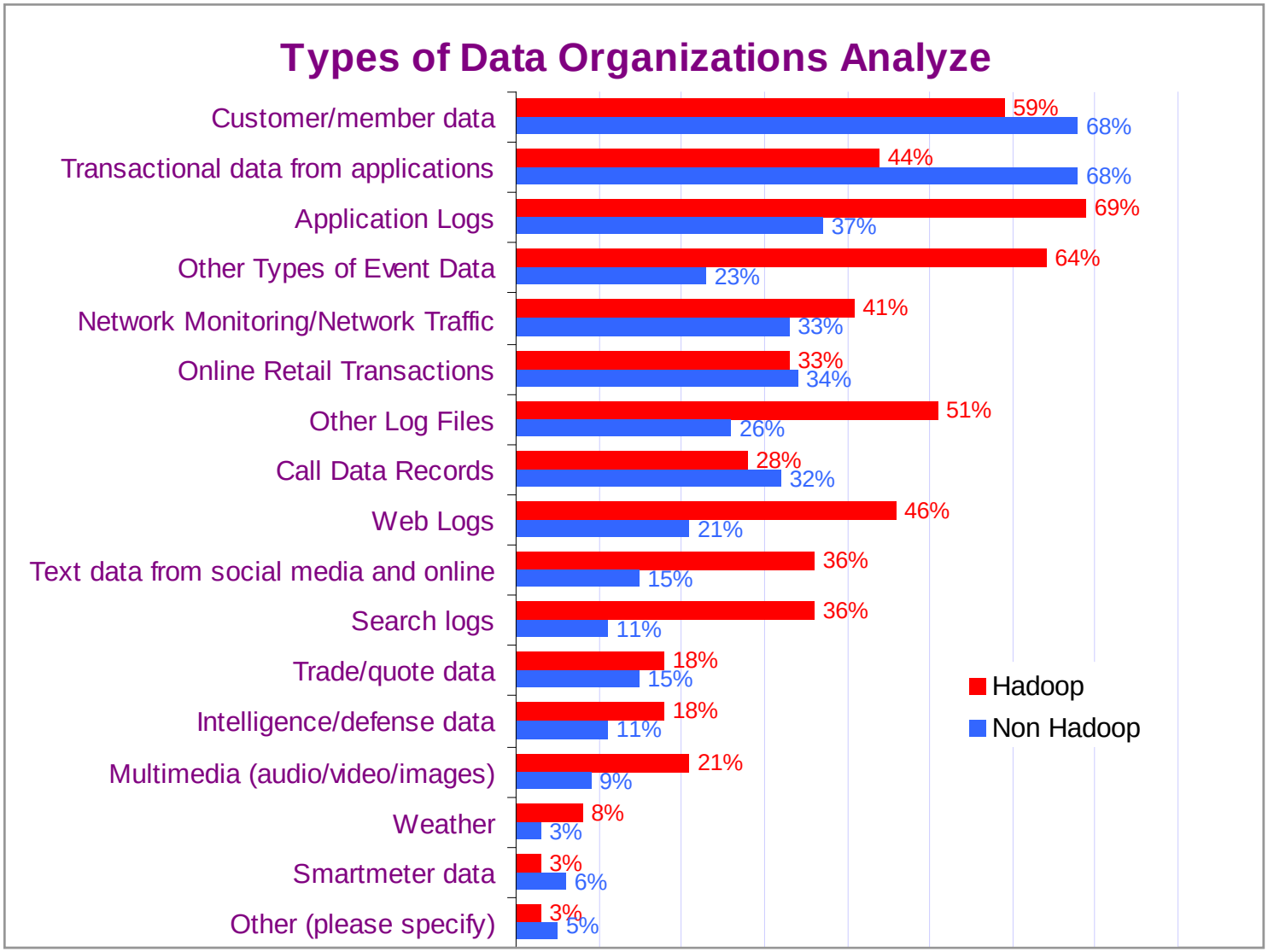


Big Data: Analytics DB Technology Impact



Big Data - Key Requirements

IMEX RESEARCH.COM



Big Data – Architectural Goals

IMEX RESEARCH.COM

Meet Requirements of V3

Volume — Optimize capital investments based on 6 Petabytes of information

Velocity — Analyze 100k records/ second to address customer satisfaction in real time

Variety — Analyze telemetry, fuel consumption, schedule and weather patterns to optimize shipping logistics.

Big Data Platform

Meet Enterprise Criterion

High availability architecture to support hardware or application failure.	Failure Tolerance
Runs on scalable hardware with the ability to dynamically add additional nodes.	Scale Economically
Security protection for granular data access control.	Security & Privacy

- Text
- Image & Video
- Acoustic
- Statistics
- Geospatial
- Mathematical
- Financial
- Times Series
- Predictive

Analyze Data in Native Format

Big Data - Market Requirements

IMEX RESEARCH.COM

Unified system: Pre-integrated for Ease of Installation and Management

- Platform – **Large Scale Indexing** Pre-integrated using Hadoop Foundation,
- Integrated Text Analytics - **Address Unstructured Data**
- **Usability** - User Friendly Admin Console including HDFS Explorer, Query Languages
- **Enterprise Class Features** – Provisioning, Storage, Scheduler, Advance Security
- Supports **search-centric, document-based XML data model**
  - store documents within a transactional repository.
- **Schema-Free**:
  - No advance knowledge of the document structure (its "schema") needed
  - Index words and values from each of the loaded documents together with its document structure.
- **Standard commodity hardware** leveraged

Big Data – Market Requirements

IMEX RESEARCH.COM

Architectural

- **Shared-nothing clustered DB architecture**
  - programmable and extensible application servers.
- Support **massive scalability** to petabytes of source data
- Support **open-source XQuery- and XSLT-driven architecture**
- Simple to Deploy, Develop and Manage (**UI & Restful Interface**)
- Support **extreme mixed workloads** - a wide variety of data types including arbitrarily hierarchical data structures, images, waveforms, data logs etc.
- Support **thousands of geographically dispersed on--line users and programs** executing variety of requests from ad hoc queries to strategic analysis
- **Loading data before declaring or discovering its structure**
- **Load data in batch and streaming fashion**
- **Integrate data from multiple sources** during load process at very high rates **Spread I/O and data across instances**
- Provide consistent **performance with linear cost**
- Leverage **Open Source SW** Lo Costs, Multiple Sources, Hadoop Foundation Tools
- Connectivity with Oracle DB, Teradata Warehouse, JDBC Connectivity,

Big Data - Market Requirements

IMEX RESEARCH.COM

Real Time Analytics Execution

- Execute **"streaming" analytic queries in real time** on incoming load data
- **Updating data in place** at full load speeds
- Scheduling and execution of **complex multi-hundred node workflows**
- Join a **billion row dimension table to a trillion row fact table** without pre-clustering the dimension table with the fact table

Performance

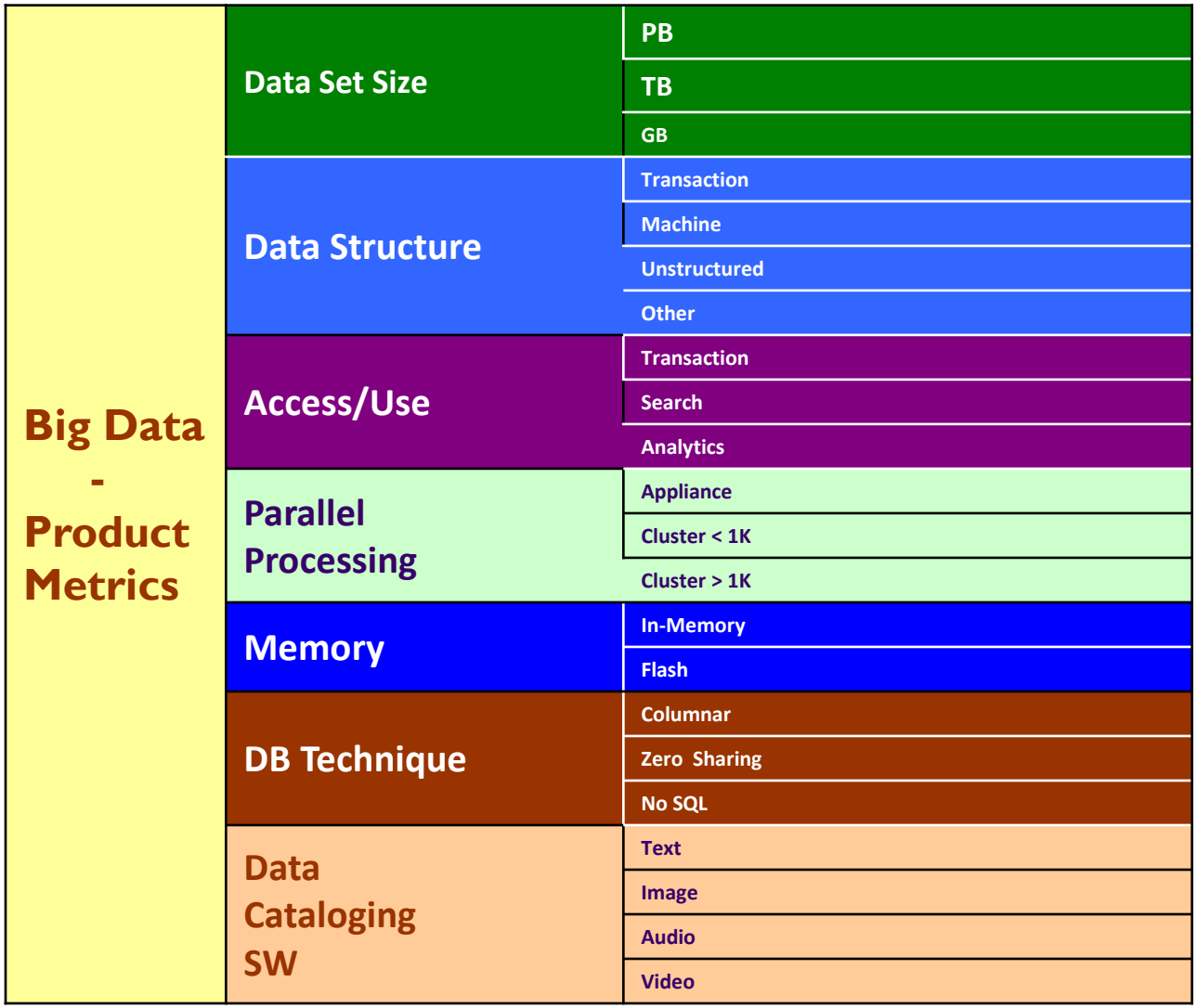
- Analyze data, at very high rates >GB/sec
- **Predictable Sub-ms response time** for highly constrained standard SQL queries

Availability

- Ability to **configure without any single point of failure**
- **Auto-Failover Extreme High Availability**
    - Automated failover and process continuation without operational interruption when processing nodes fail

Big Data – Product Metrics Choices

IMEX RESEARCH.COM



Advantage: Big Data Products

IMEX RESEARCH.COM

Characteristic	Legacy Paradigm	Big Data Paradigm
Structure	•Transactional/Corporate	•Unstructured/Derivative/Internet
Mode	•Data Collection	•Data Analysis
Focus	•Find Answers	•Find Questions
Facility	•Reportive / What Happened?	•Analytic / Why did it Happen? Predictive / What will Happen Next?
Opportunity	•Very Small Growth	•Massive Growth
Players	•Legacy Players	•Agile Start Ups, well funded
Impact	•Analyze Existing Businesses	•Create New Businesses

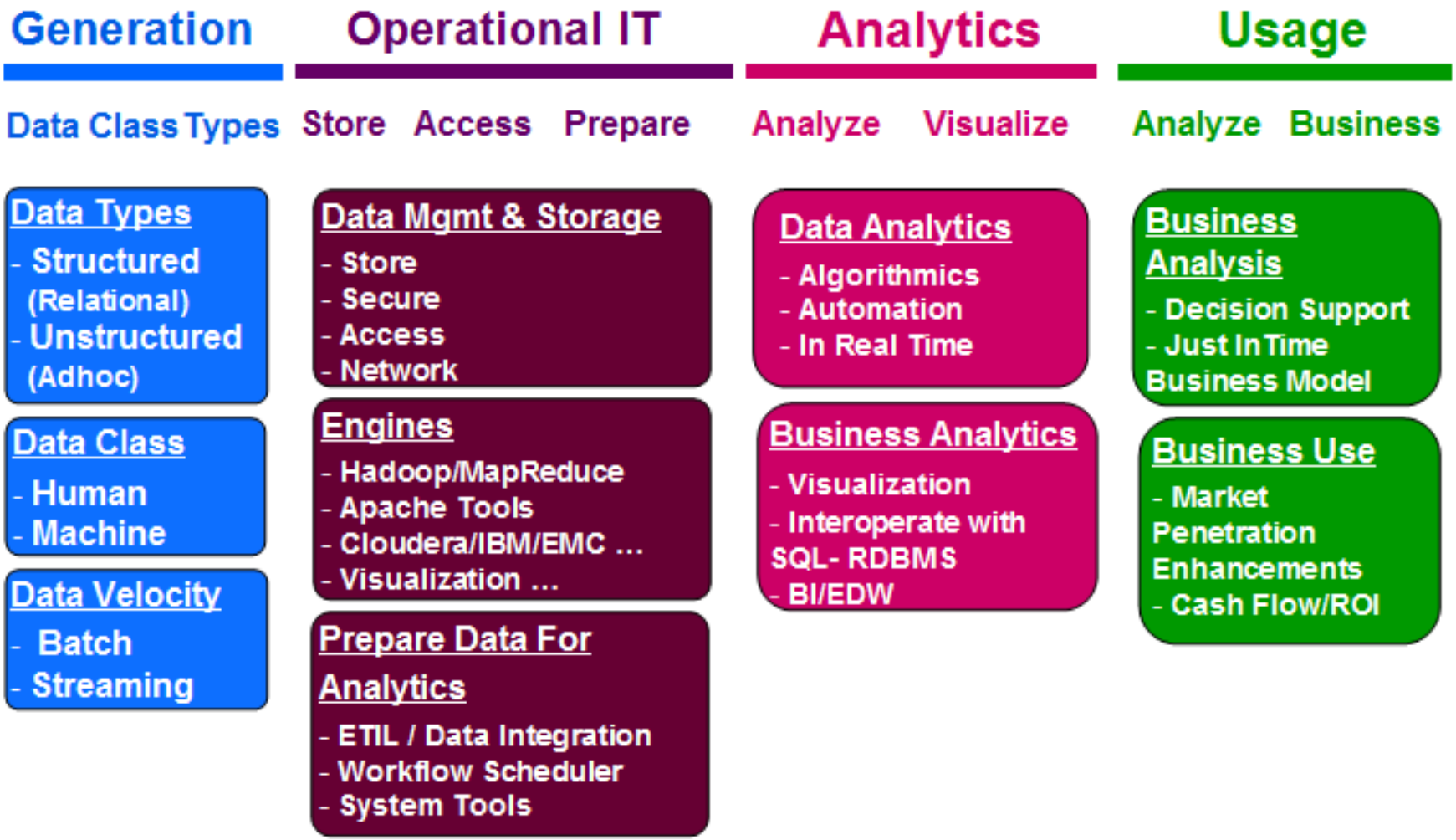
Advantage: Big Data Products

IMEX RESEARCH.COM

Characteristic	Traditional RDBMS	Big Data/MapReduce
Data Size	•GB	•PB
Access	•Interactive	•Batch/Near Real-Time
Latency	•Low	•High
Data Updates	•Read & Write Many Times	•Write Once Read Many Times
Schema/Structure	•Static Schema	•Dynamic Schema
Language	•SQL	•UQL/Procedural (Java,C++..)
Integrity	•High	•Not 100%
Works Well for	•Process Intensive Jobs	•Data Intensive Jobs
Works Well w Data Size	•Gigabytes	•Petabytes
Data/Processing Interactions	•Low Latency/High BW – precursor to success. Ntwk. BW can be a bottleneck causing nodes to be idle	•Sends Code to Data, instead of Sending Data to other Nodes (Requiring Lower BW in Cluster)
Fault Tolerance	•Coordinating Processes with Node Failures – a challenge	•Fault Tolerant for HW/SW Failures
Access	•Interactive	•Batch/Near Real-time
Scaling	•Non-linear	•Linear
Pgm-Distribution of Jobs	•Difficult	•Simple & Effective

Big Data Ecosystem

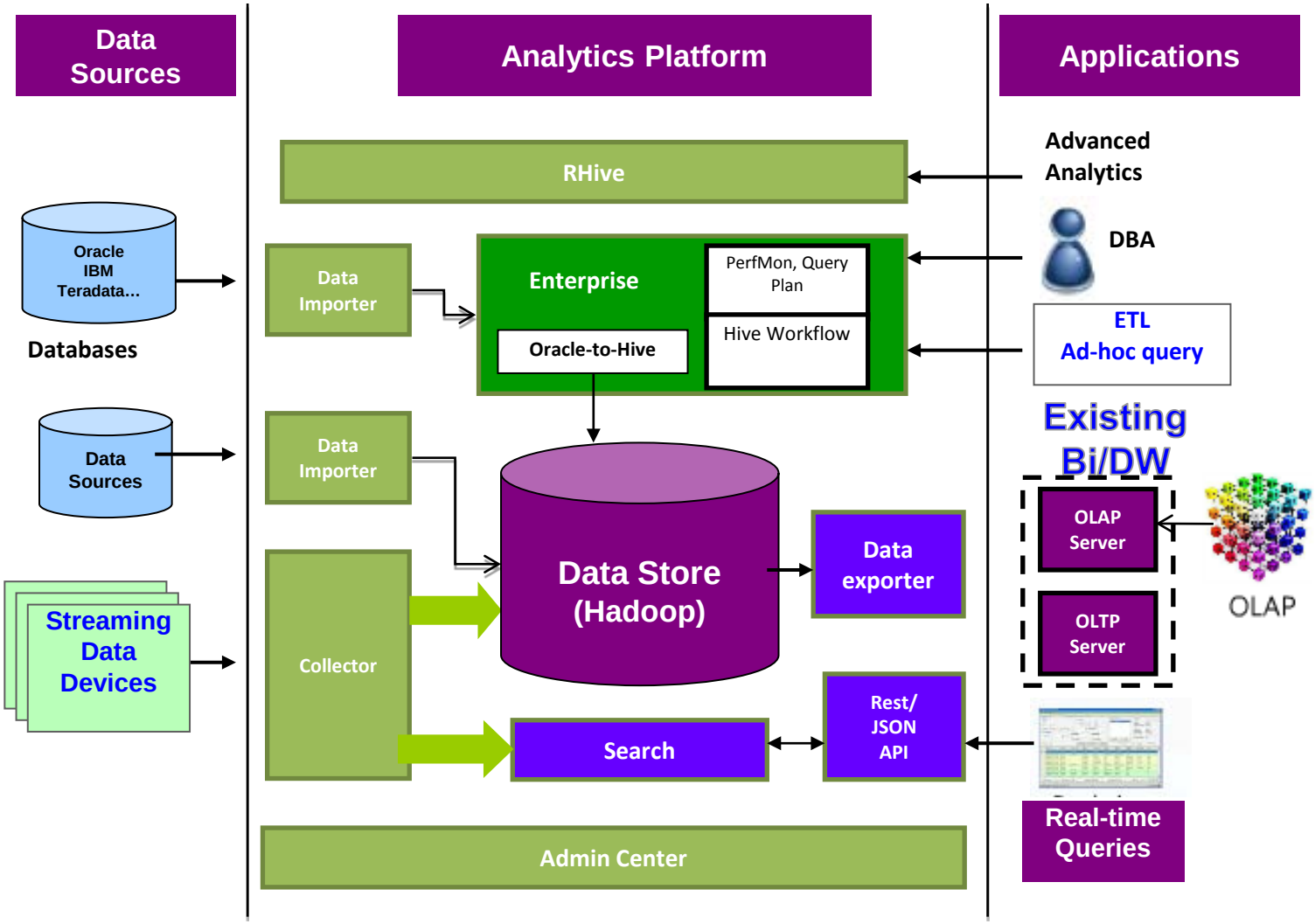
IMEX RESEARCH.COM



Big Data Stack

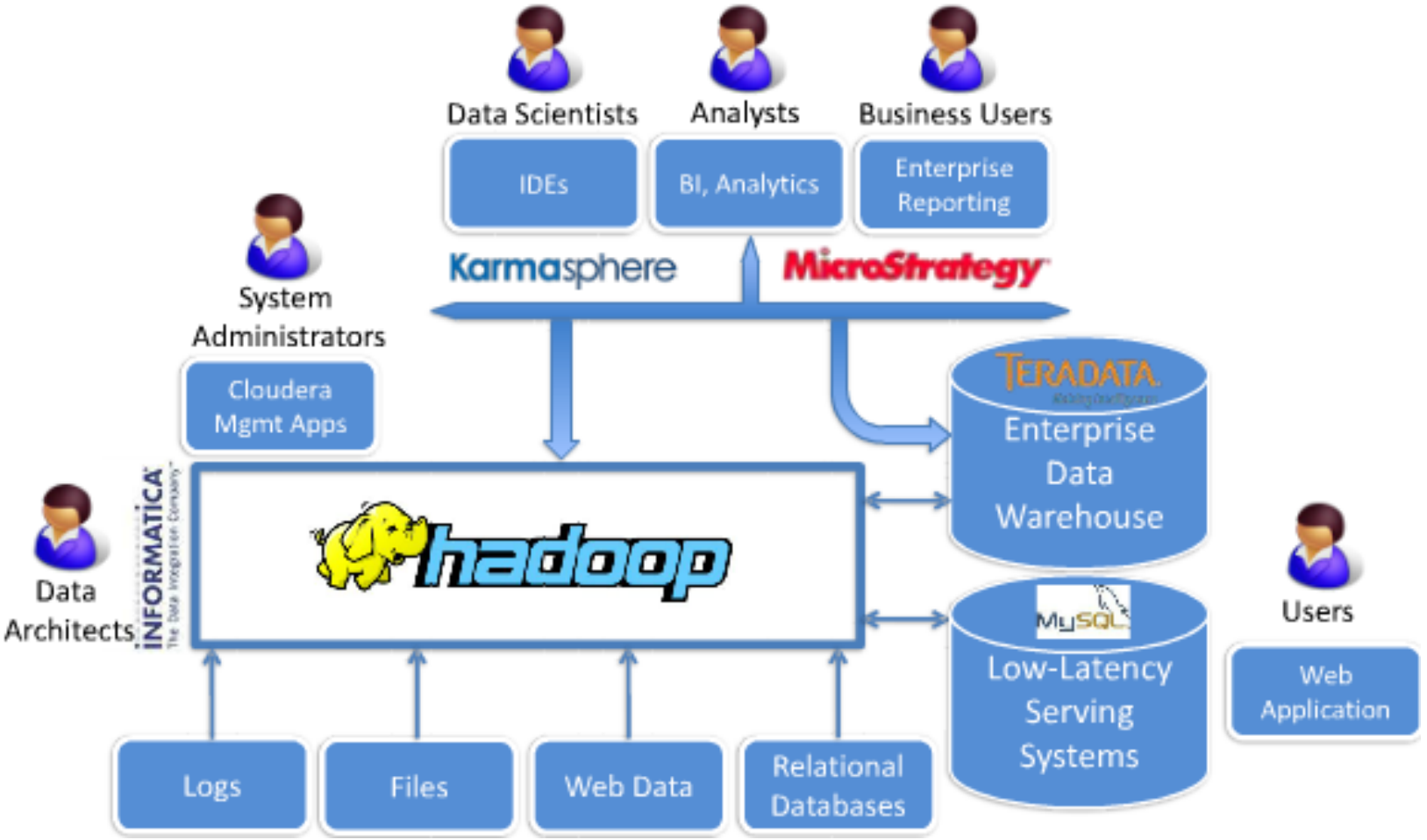
IMEX RESEARCH.COM

Merging Hadoop innovations into Nextgen DBMS



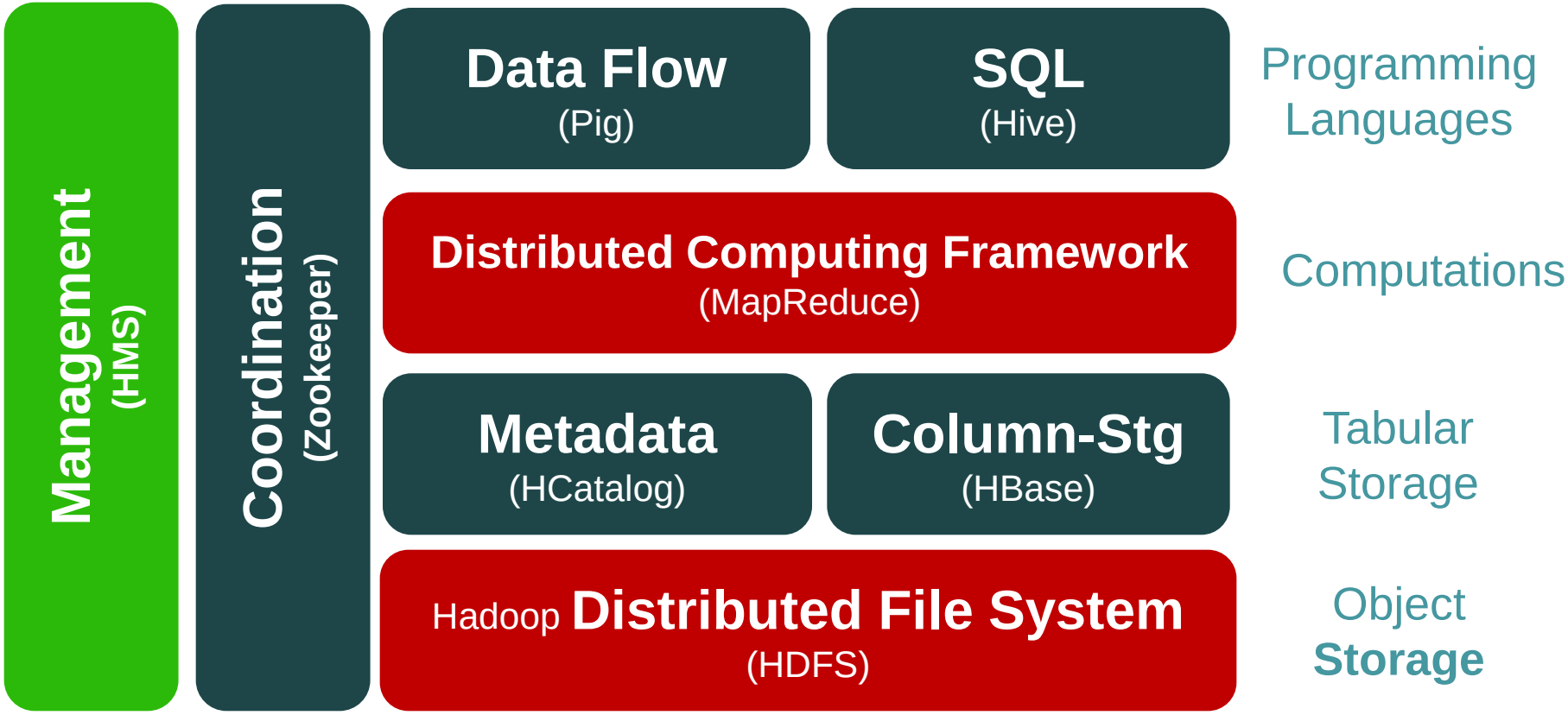
Hadoop's Fit in Enterprise Stack

IMEX RESEARCH.COM



Big Data - Hadoop Architecture

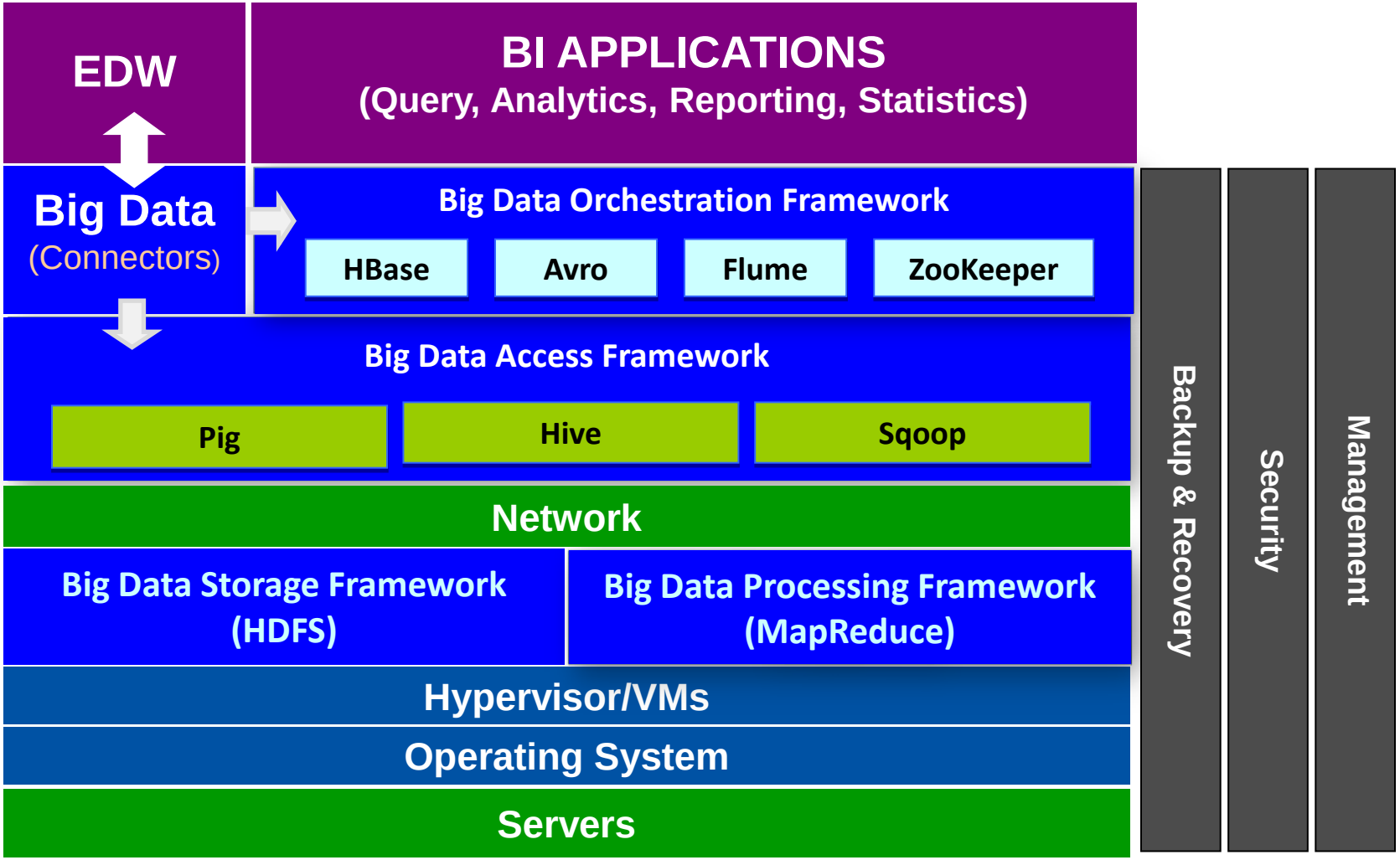
IMEX RESEARCH.COM



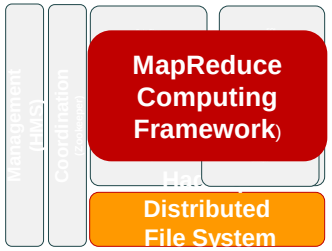
Big Data Connectors to EDW/BI

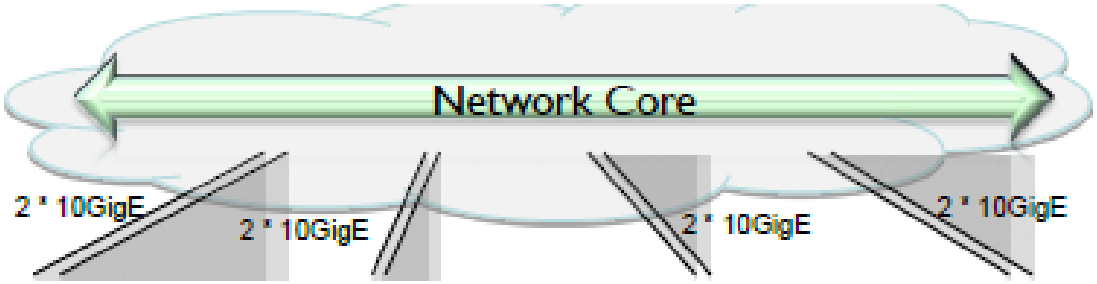
IMEX RESEARCH.COM

BI Framework - Interoperable with Enterprise Data Warehousing

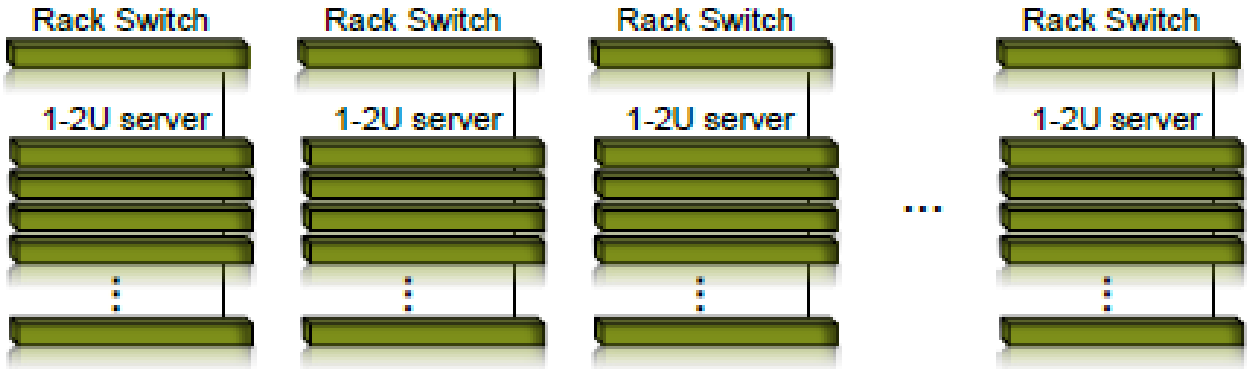


Big Data Infrastructure – Map Reduce



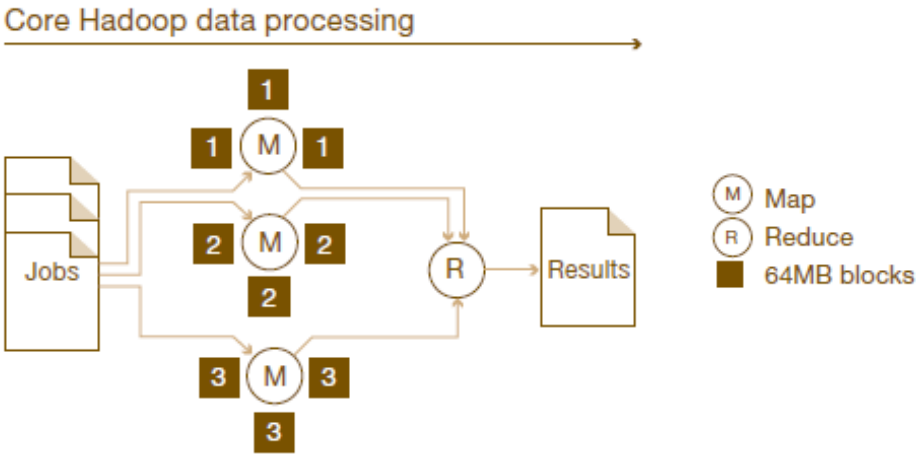


- 20-40 nodes / rack
- 16 Cores
- 48G RAM
- 6-12 * 2TB disk
- 1-2 GigE to node



Map Reduce

- A Distributed Computing Model
- Typical Pipeline:
  Input>**Map**>Shuffle/Sort>**Reduce**>Output
- Easy to Use , Developer writes few functions,
  Moves compute to Data
- Schedules work on HDFS node with data
- Scans through data, reducing seeks
- Automatic Reliability and re-execution on failure

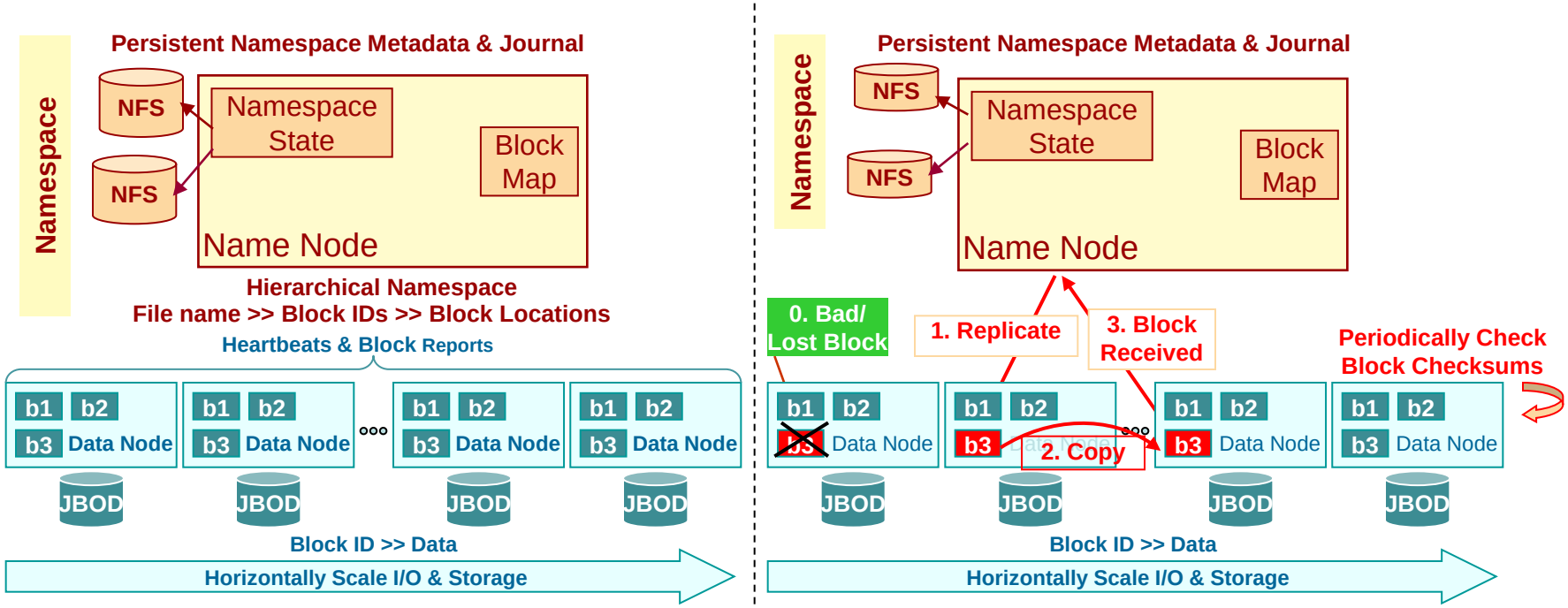


Big Data Infrastructure – HDFS

IMEX RESEARCH.COM

HDFS Architecture

Actively Maintaining High Availability



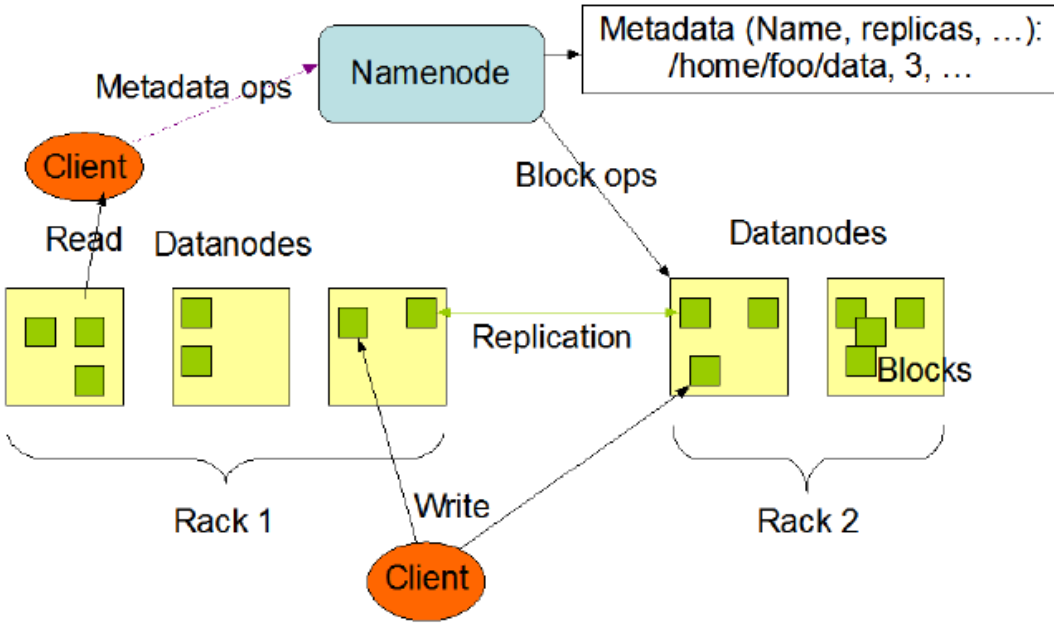
HDFS

- Immutable File System – Read, Write, Sync/Flush – No random writes
- Storage Server used for Computation – Move Computation to Data
- Fault Tolerant & Easy Management – Built In Redundancy, Tolerates Disk & Node Failure, Auto-Managing addition/removal of nodes, One operator/8K nodes
- Not a SAN but high bandwidth network access to data via Ethernet
- Used typically to Solve problems not feasible with traditional systems: Large Storage Capacity >100PB raw, Large IO/computational BW >4K node/cluster, scale by adding commodity HW, Cost ~$1.5/GB incl. MR cluster

Hadoop Distributed File System

IMEX RESEARCH.COM

HDFS Architecture



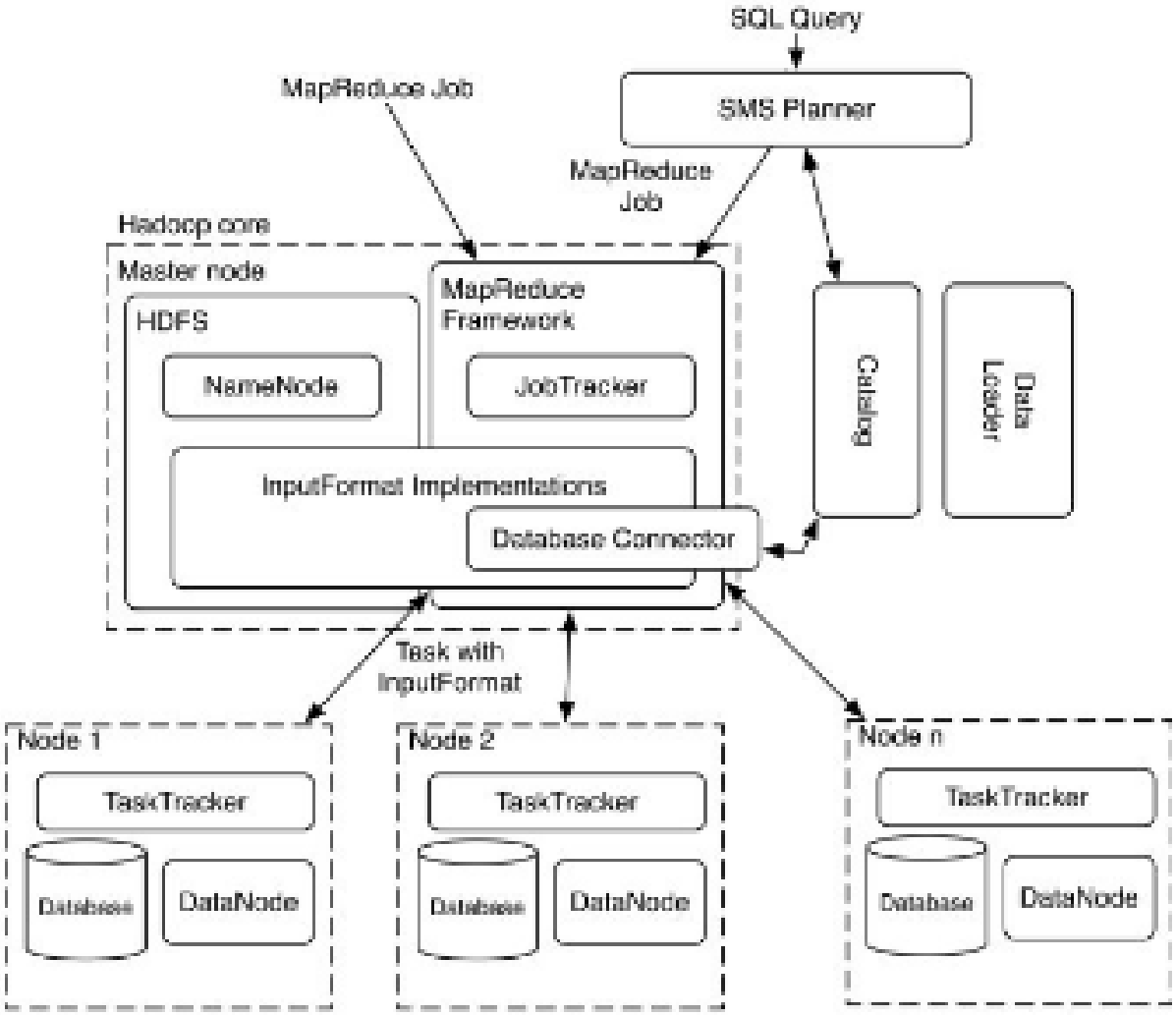
HDFS Characteristics

- **Based on Google GFS** (Google File System)
- **Redundant Storage** for massive amounts of data
- **Data is distributed across all nodes** at load time – efficient MapReduce processing
- **Runs on commodity hardware** – assumes high failure rate for components
- **Works well with lots of large files**
- **Built around Write once** – Read many times"
- **Large Streaming Reads** – Not random access
- **High Throughtput** more important than low latency

Hadoop Architecture - Overview

IMEX RESEARCH.COM

Hadoop Data Processing Architecture



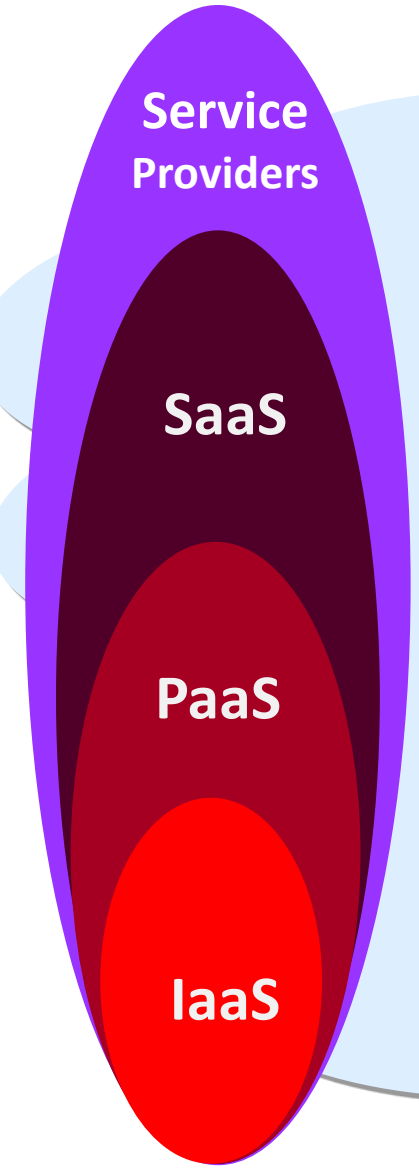
IMEX RESEARCH.COM

Key Technologies Required for Big Data

- **Cloud Infrastructure**
- **Virtualization**
- **Networking**
- **Storage**
  - In-Memory Data Base (Solid State Memory)
  - Tiered Storage Software (Performance Enhancement)
  - Deduplication (Cost Reduction)
  - Data Protection (Back Up, Archive & Recovery)

Cloud Infrastructure for Big Data

IMEX RESEARCH.COM



Cloud Services Providers
Public – Mutitenancy,OnDemand
Private - On Premises, Enterprise
Hybrid – Interoperable P2P

Examples
Public - BT, Telstra, T-Systems France Telecom
Private –
Hybrid – IBM/Cloudburst,

Software-as-a-Service
- Servers, Network, Storage
- Management, Reporting

Examples
eMail - Yahoo!,Google…
Collaboration - Facebook,Twitter …
Bus.Apps - SalesForce, GoogleApps, Intuit…

Platform Tools & Services
- Deploy developed platforms ready for Application SW on Cloud Aware Infrastructure

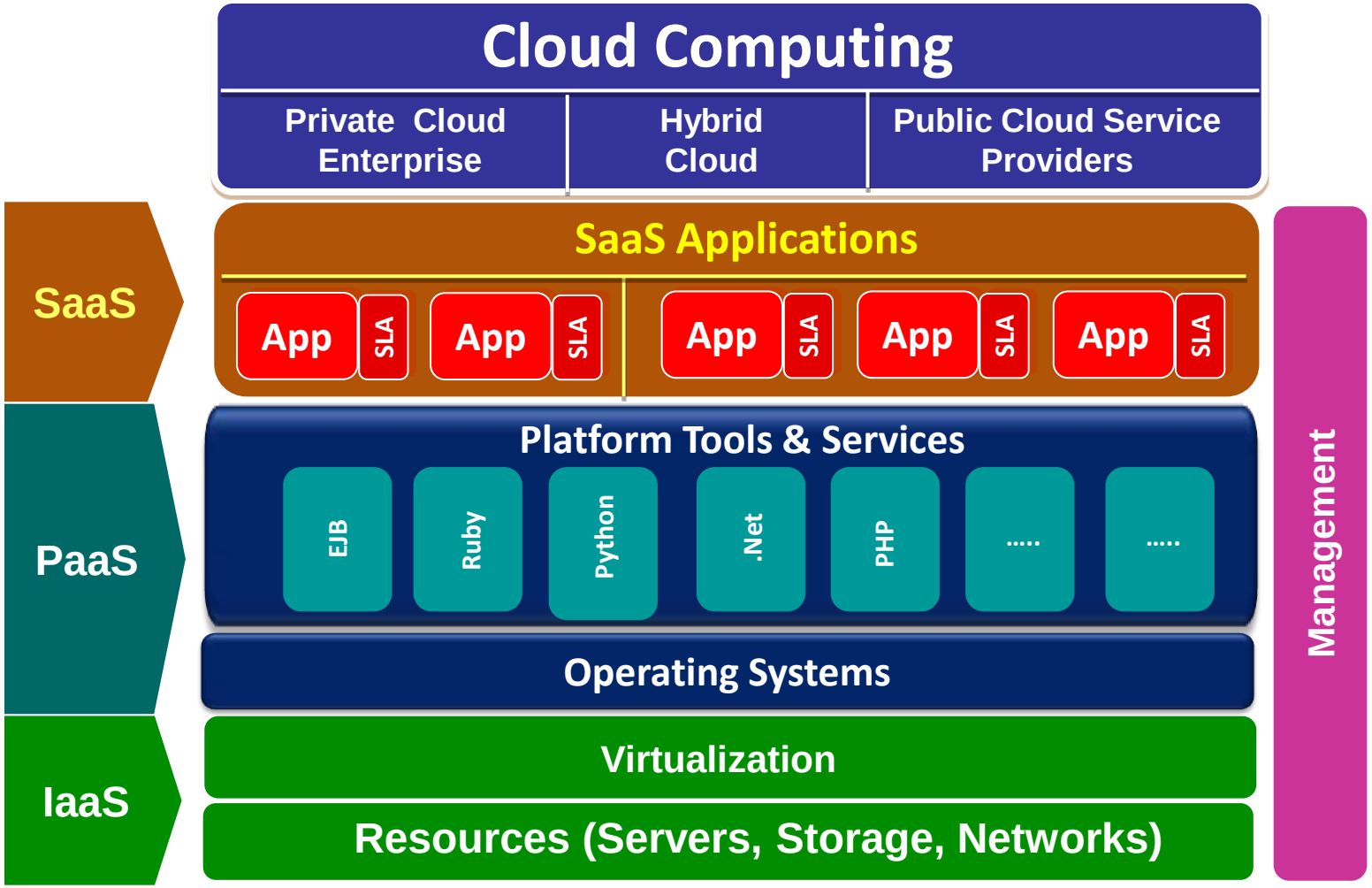
Examples
Amazon EC2
Force.com
Navitaire

Infrastructure HW & Services
- Servers, Network, Storage
- Management, Reporting

Examples
Amazon S3
Nirvanix

Cloud Infrastructure for Big Data

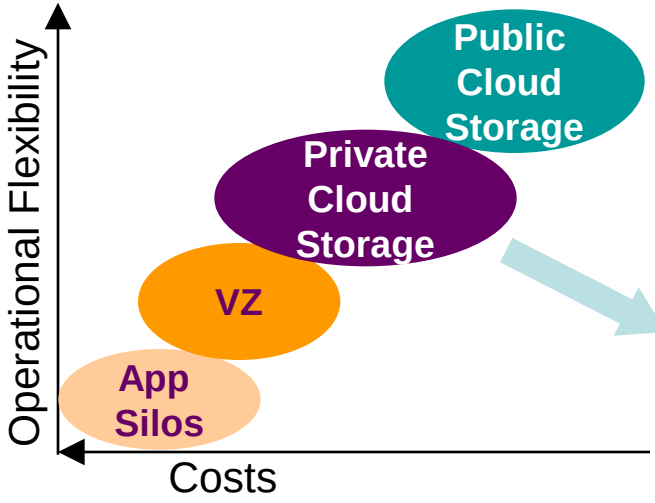
IMEX RESEARCH.COM



Application's SLA dictates the Resources Required to meet specific requirements of Availability, Performance, Cost, Security, Manageability etc.

Private Cloud Requirements for Big Data

IMEX RESEARCH.COM



Service Catalog

- Choose pre-defined IT Services/user/dept.
- Define SLA to efficiently meet services

Self-Service

- Access Resources on demand to speed deployment and delivery
- Scale Resource Up/down
- to optimize their usage,
- Release when not needed

Private Cloud Storage

Service Analytics

- Monitor & Analyze Usage for Charge back
- Interactively auto-tune performance

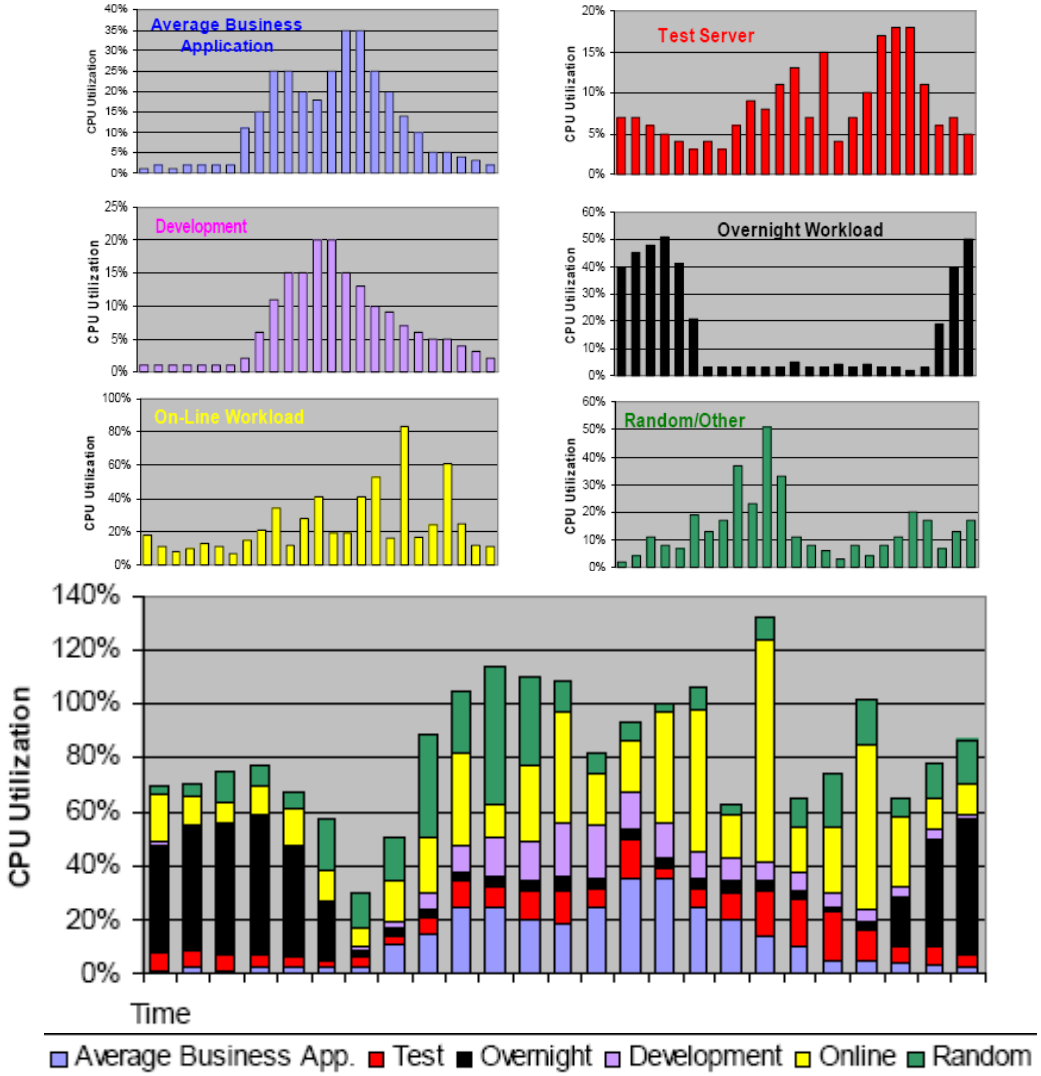
Availability with SLA requirements

Automation

- Automated Provisioning
- Moving Data & Processes Seamlessly
- Allocation, Self Tuning of Resources to meet Workload Requirements

Virtualization: Workloads Consolidation

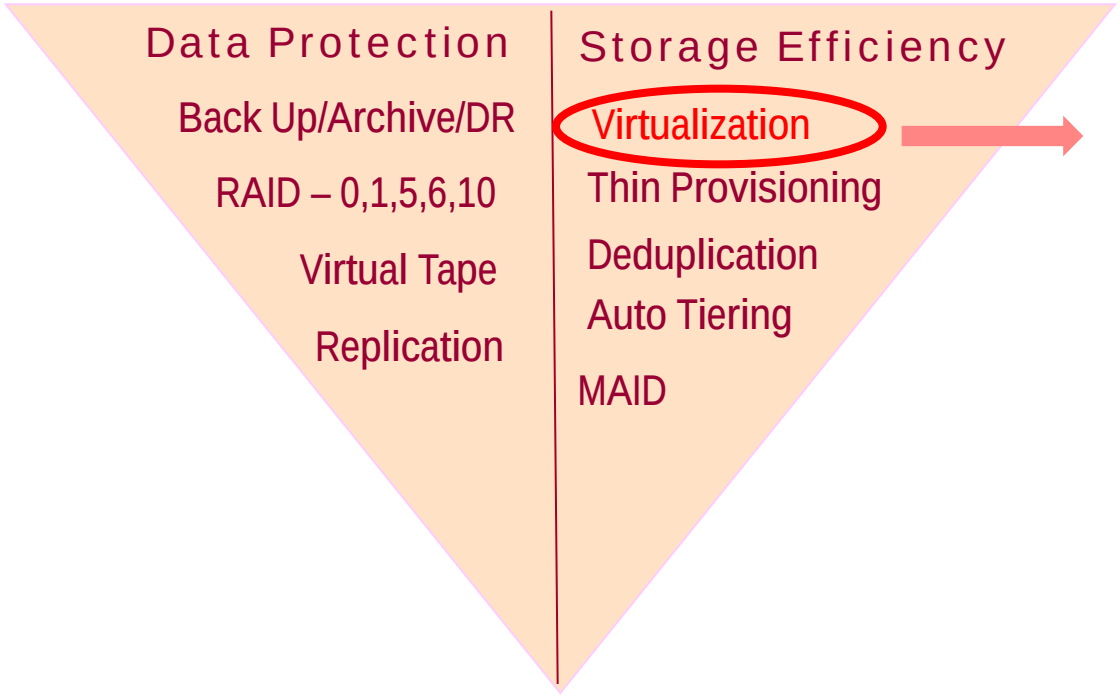
IMEX RESEARCH.COM



Source: Dan Olds & IMEX Research 2009

- A single server 1.5x larger than standard 2-way server will handle consolidated load of 6 servers.
- VZ manages the workloads + important apps get the compute resources they need automatically w/o operator intervention.
- Physical consolidation of 15-20:1 is easily possible
- Reasonable goal for VZ x86 servers – 40-50% utilization on large systems (>4way), rising as dual/quad core processors becomes available
- Savings result in Real Estate, Power & Cooling, High Availability, Hardware, Management

Storage Architecture - Impact from VZ



Virtualization (VZ) requires Shared Storage for
- VMotion
- Storage VMotion
- HA/DRS
- Fault Tolerance

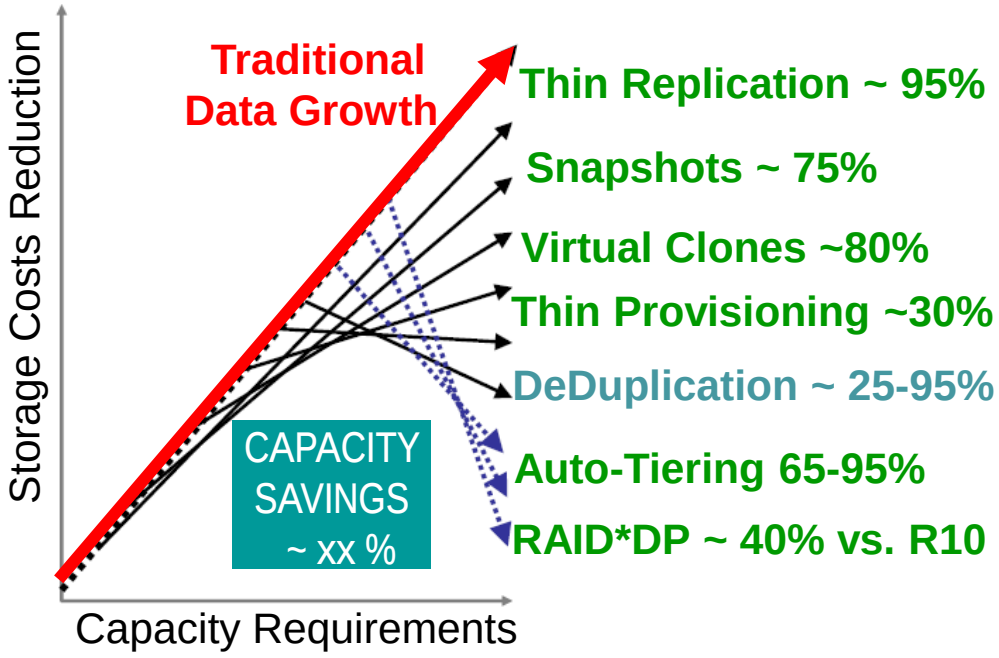
Additional Capacity Consumed for
- VZ snapshots,
- VM Kernel etc.

Source: IMEX Research SSD Industry Report ©2011

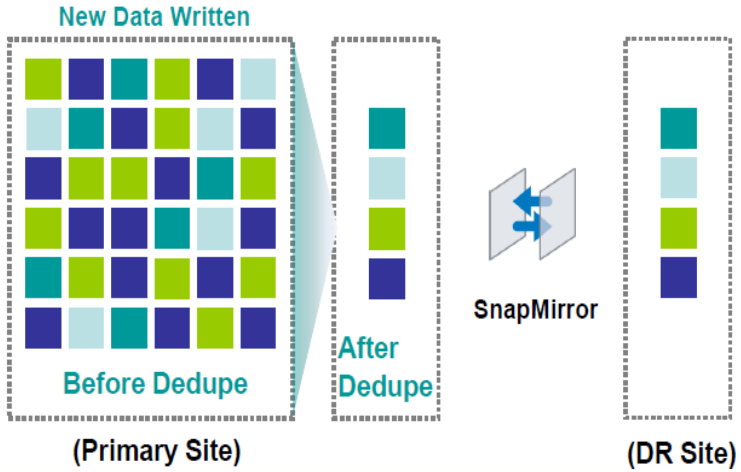
Storage – Issues & Solutions

IMEX RESEARCH.COM

Technologies Reducing Storage Costs



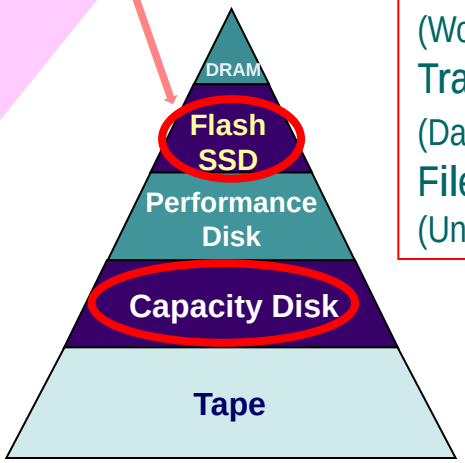
- Replicates only the deduplicated blocks
- Only unique data is replicated to the DR site



Storage Architecture Impacting Big Data

IMEX RESEARCH.COM

Data Protection	Storage Efficiency
Back Up/Archive/DR	Storage Virtualization
RAID – 0,1,5,6,10	Thin Provisioning
Virtual Tape	Deduplication
Replication	Auto-Tiering
	MAID



Auto-Tiering System using Flash SSDs

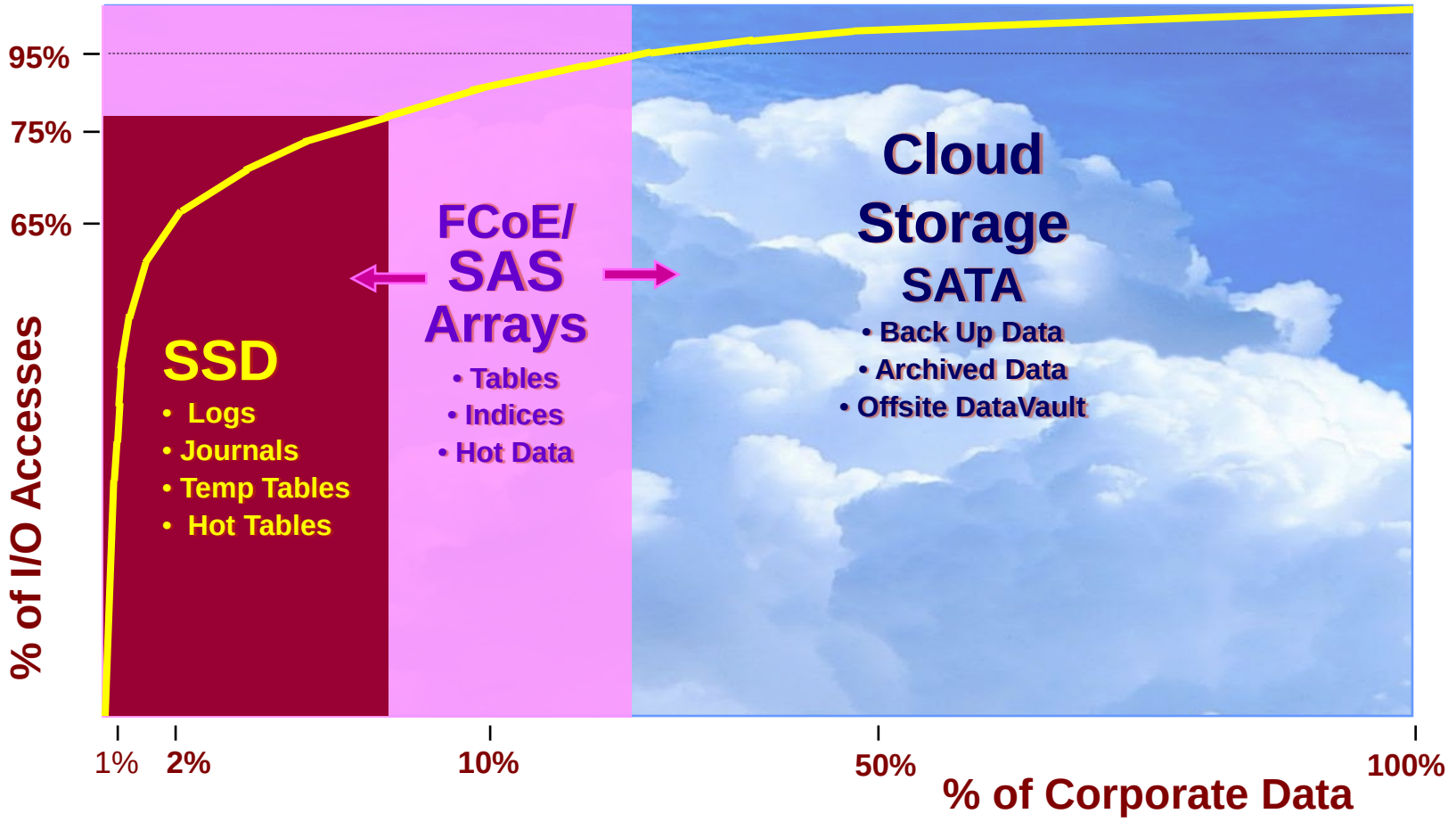
- Data Class
  (Tiers 0,1,2,3)
- Storage Media Type
  (Flash/Disk/Tape)
- Policy Engines
  (Workload Mgmt.)
- Transparent Migration
  (Data Placement)
- File Virtualization
  (Uninterrupted App.Opns.in Migration)

Source: IMEX Research SSD Industry Report ©2011

Data Storage: Hierarchical Usage

IMEX RESEARCH.COM

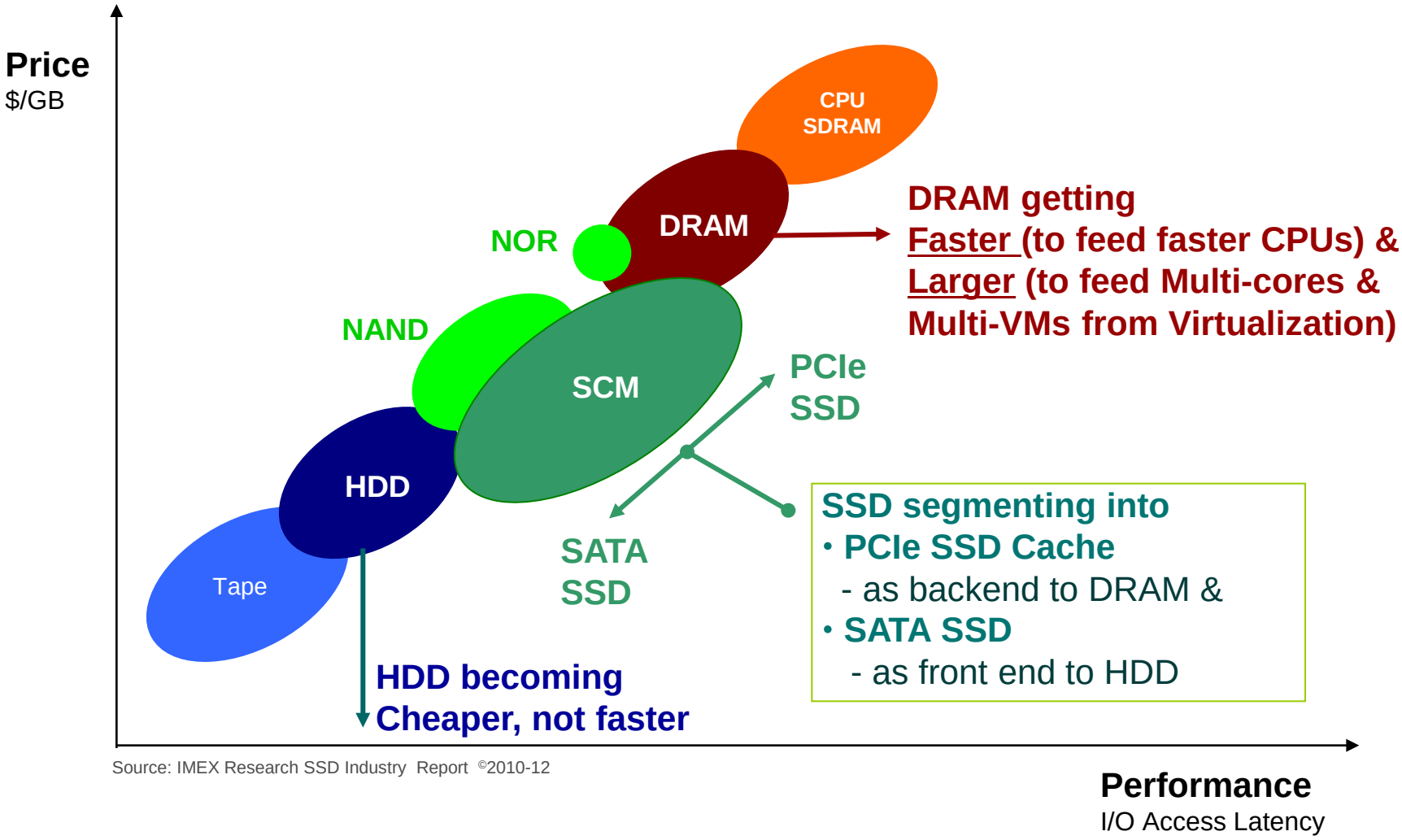
I/O Access Frequency vs. Percent of Corporate Data



Source:: IMEX Research - Cloud Infrastructure Report ©2009-12

SSD Storage: Filling Price/Perf.Gaps

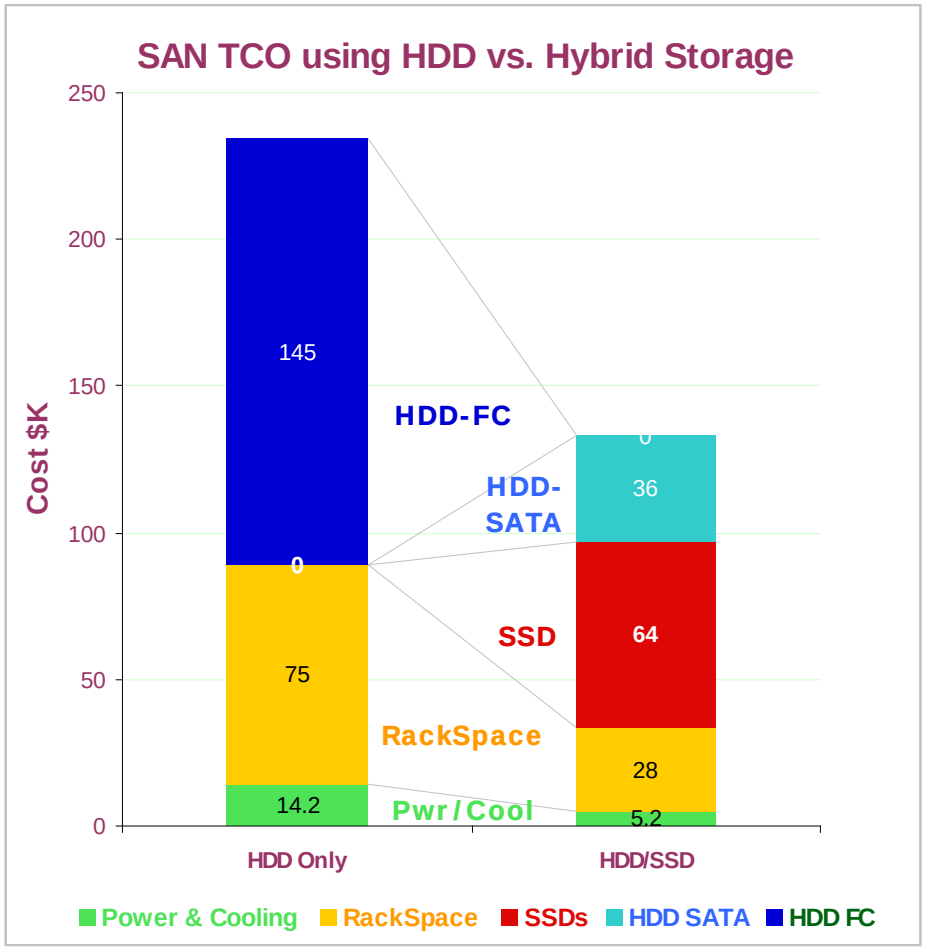
IMEX RESEARCH.COM

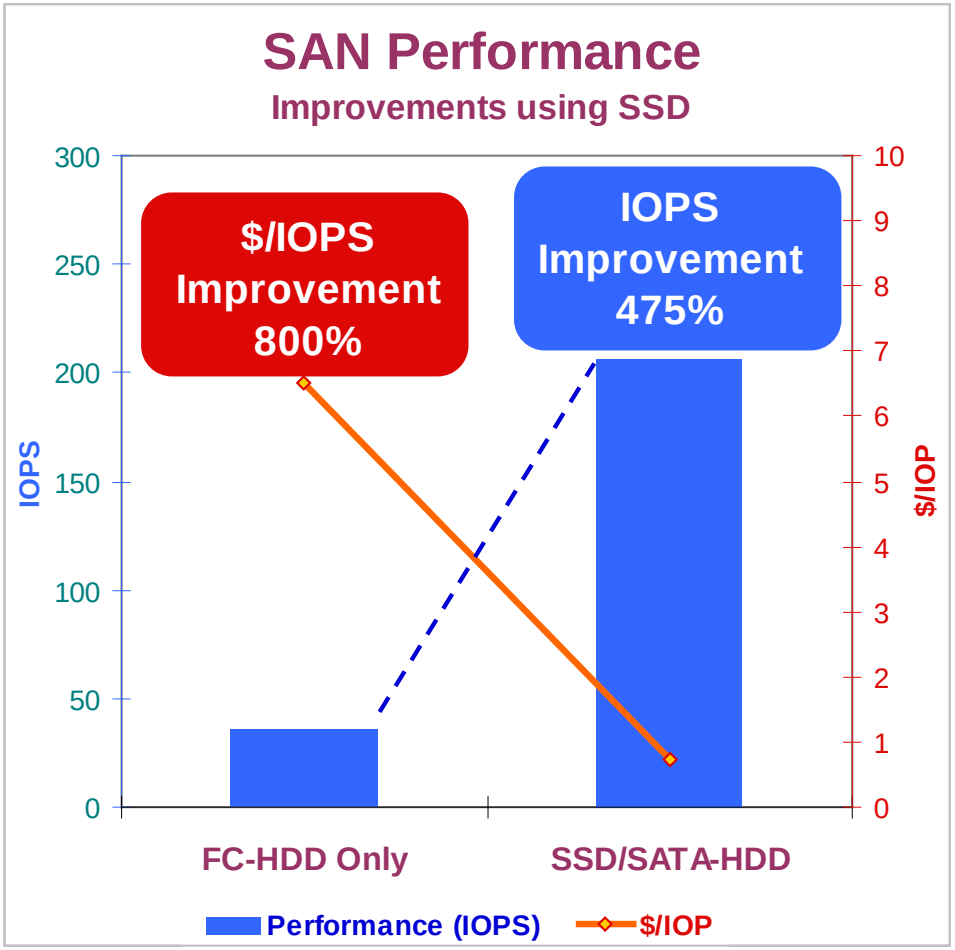


Source: IMEX Research SSD Industry Report ©2010-12

SSD Storage - Performance & TCO

IMEX RESEARCH.COM

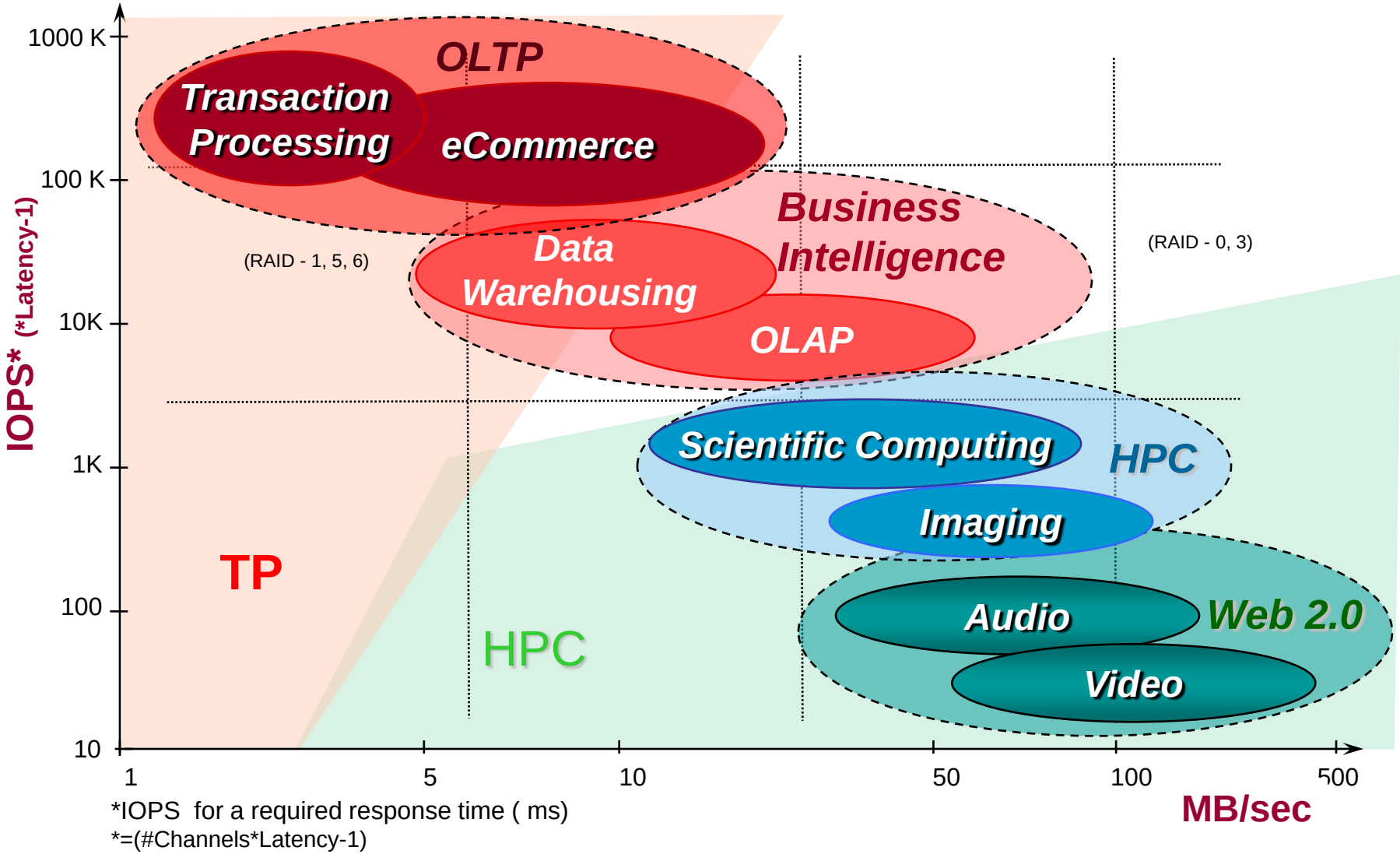




Source: IMEX Research SSD Industry Report ©2011

Workloads Characterization

IMEX RESEARCH.COM



*IOPS for a required response time (ms)

*=(#Channels*Latency-1)

Source:: IMEX Research - Cloud Infrastructure Report ©2009-12

Workloads Characterization

IMEX RESEARCH.COM

Storage performance, management and costs are big issues in running Databases

- **Data Warehousing Workloads are I/O intensive**
  - Predominantly read based with low hit ratios on buffer pools
  - High concurrent sequential and random read levels
    - ✓ Sequential Reads requires high level of I/O Bandwidth (MB/sec)
    - ✓ Random Reads require high IOPS)
  - Write rates driven by life cycle management and sort operations
- **OLTP Workloads are strongly random I/O intensive**
  - Random I/O is more dominant
    - ✓ Read/write ratios of 80/20 are most common but can be 50/50
    - ✓ Can be difficult to build out test systems with sufficient I/O characteristics
- **Batch Workloads are more write intensive**
  - Sequential Writes requires high level of I/O Bandwidth (MB/sec)
- **Backup & Recovery times are critical for these workloads**
  - Backup operations drive high level of sequential IO
  - Recovery operation drives high levels of random I/O

Source: IMEX Research SSD Industry Report ©2011

Best Practices – Storage in Big Data Apps

IMEX RESEARCH.COM

Goals & Implementation

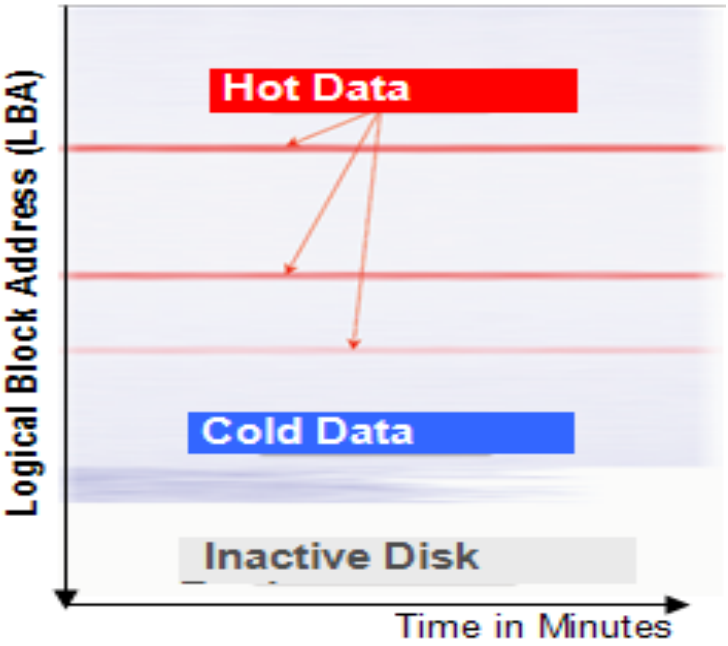
- Establish **Goals for SLAs** (Performance/Cost/Availability), BC/DR (RPO/RTO) & Compliance
- **Increase Performance for DB, OLTP and OLAP Apps**:
  - Random I/O > 20x , Sequential I/O Bandwidth > 5x
  - Remove Stale data from Production Resources to improve performance
- Use **Partitioning Software to Classify Data**
  - By Frequency of Access (Recent Usage) and
  - Capacity (by percent of total Data) using general guidelines as:
  - Hyperactive (1%), Active (5%), Less Active (20%), Historical (74%)

Implementation

- **Optimize Tiering** by Classifying Hot & Cold Data
  - Improve Query Performance by reducing number of I/Os
  - Reduce number of Disks Needed by 25-50% using advance compression software achieving 2-4x compression
- **Match Data Classification vs.Tiered Devices** accordingly
  - Flash, High Perf Disk, Low Cost Capacity Disk, Online Lowest Cost Archival Disk/Tape
- **Balance Cost vs. Performance** of Flash
  - More Data in Flash > Higher Cache Hit Ratio > Improved Data Performance
- **Create and Auto-Manage Tiering** (Monitoring, Migrations, Placements) without manual intervention

Best Practices: I/O Forensics in Storage-Tiering

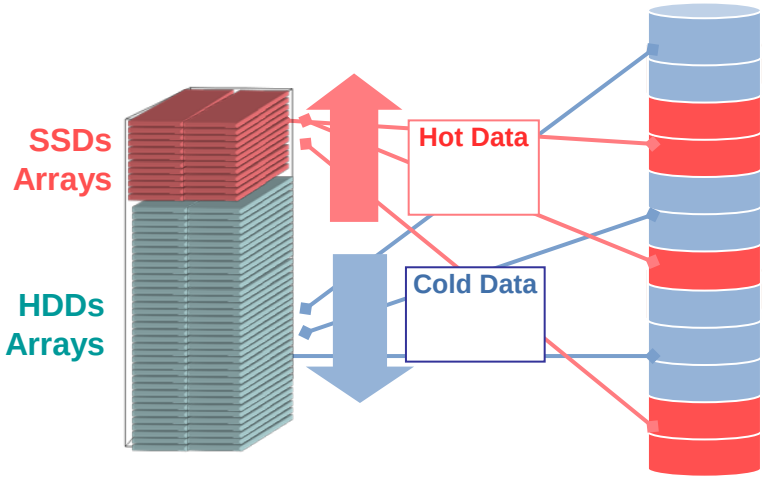
IMEX RESEARCH.COM



Storage-Tiered Virtualization

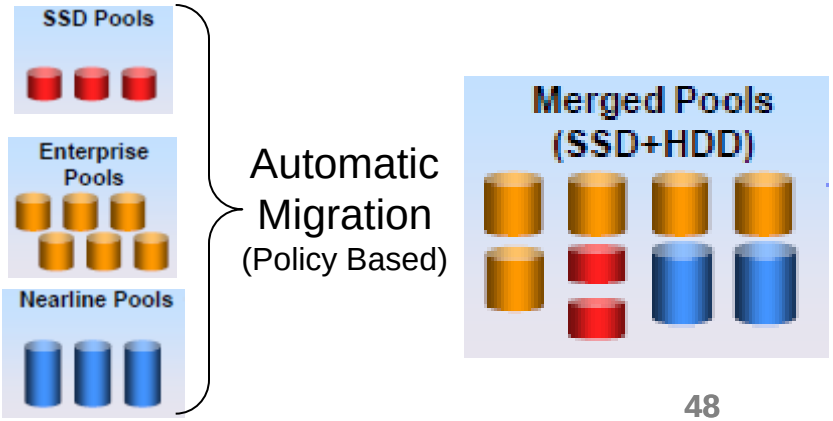
Storage-Tiering at LBA/Sub-LUN Level

Physical Storage **Logical Volume**



LBA Monitoring and Tiered Placement

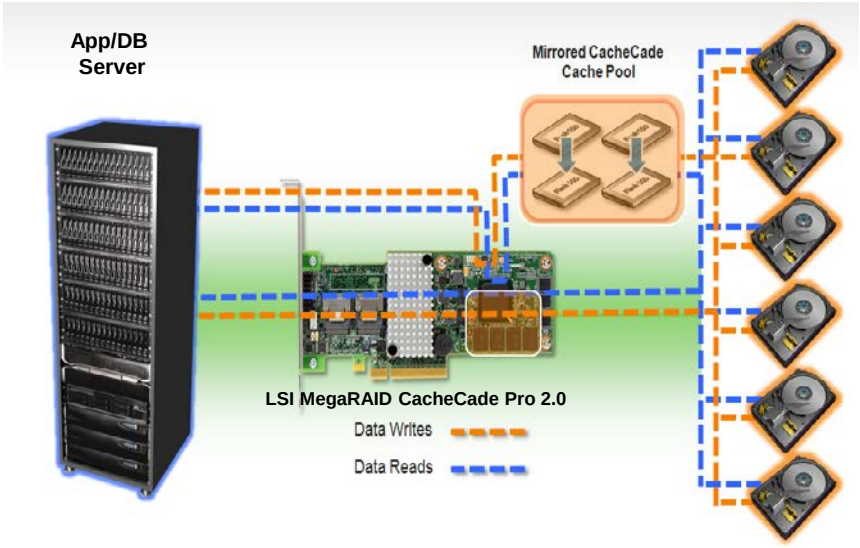
- Every workload has unique I/O access signature
- Historical performance data for a LUN can identify performance skews & hot data regions by LBAs



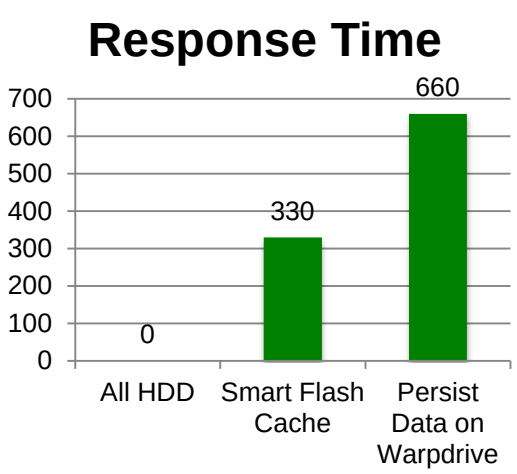
Source: IBM & IMEX Research SSD Industry Report 2011 ©IMEX 2010-12

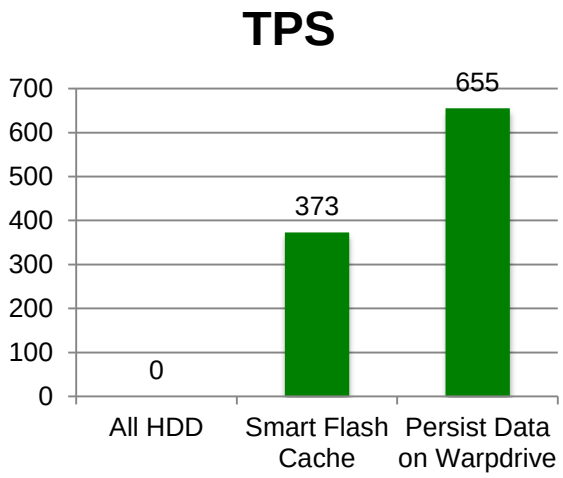
Best Practices: Cached Storage

IMEX RESEARCH.COM



Application	Improvement over Cached vs.HDD only
Oracle OLTP Benchmarks	681%
SQL Server OLTP Benchmark	1251%
Neoload (Web Server Simulation	533%
SysBench (MySQL OLTP Server)	150%



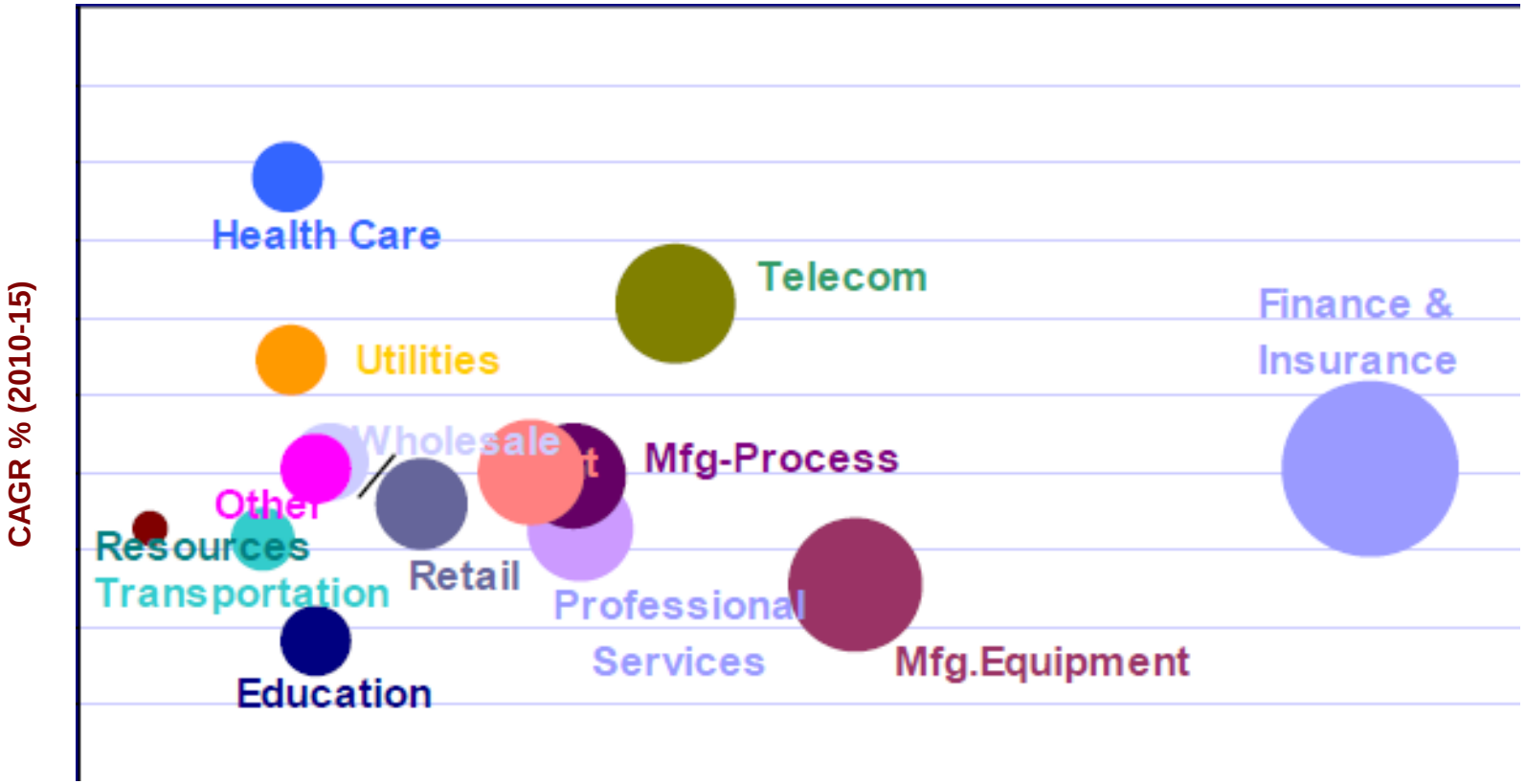


Big Data Targets: Analytics

IMEX RESEARCH.COM

Key Industries Benefitting from Big Data Analytics

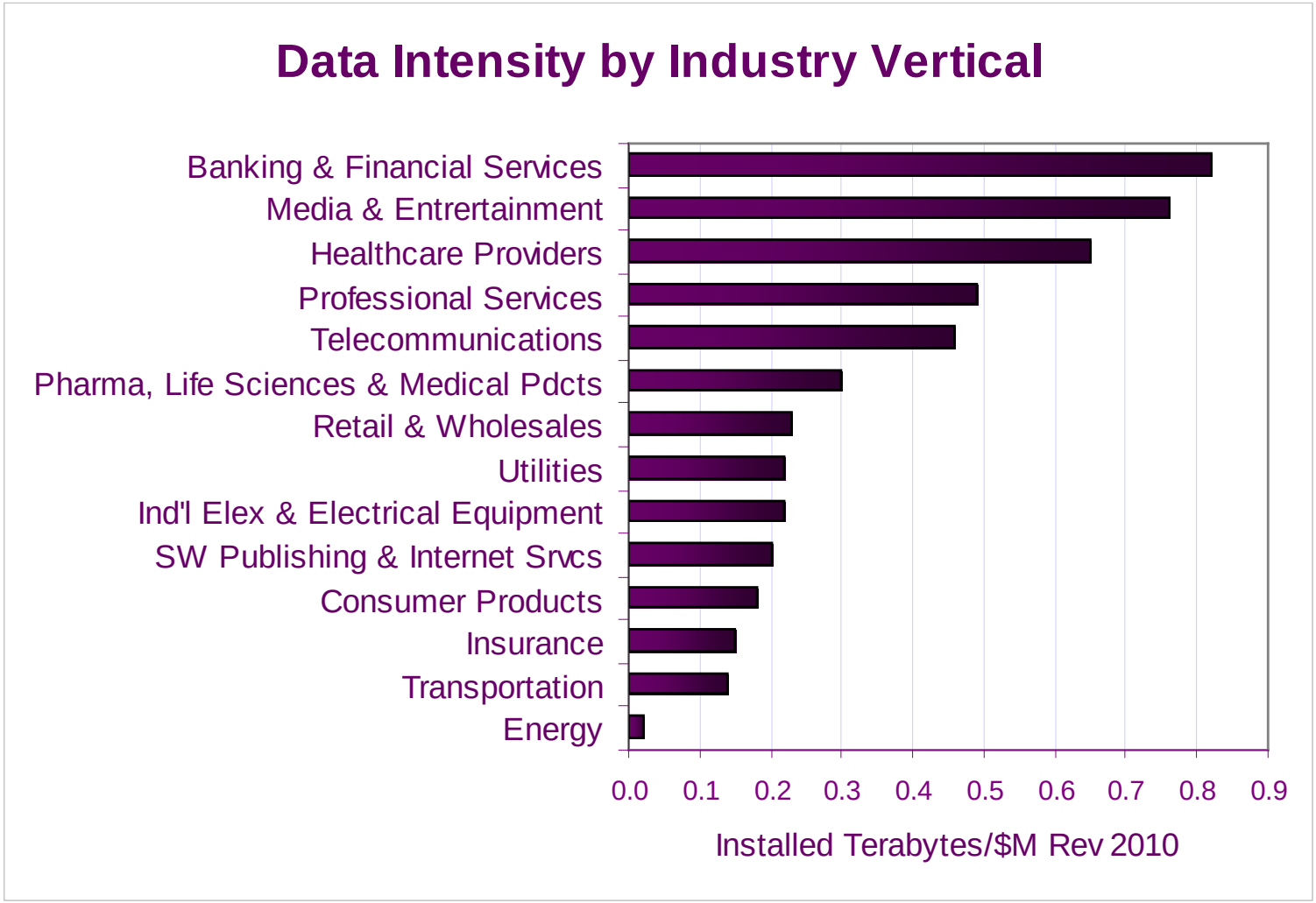
Global IT Spending by Industry Verticals 2010-15 $B



Big Data Targets – Storage Infrastructure

IMEX RESEARCH.COM

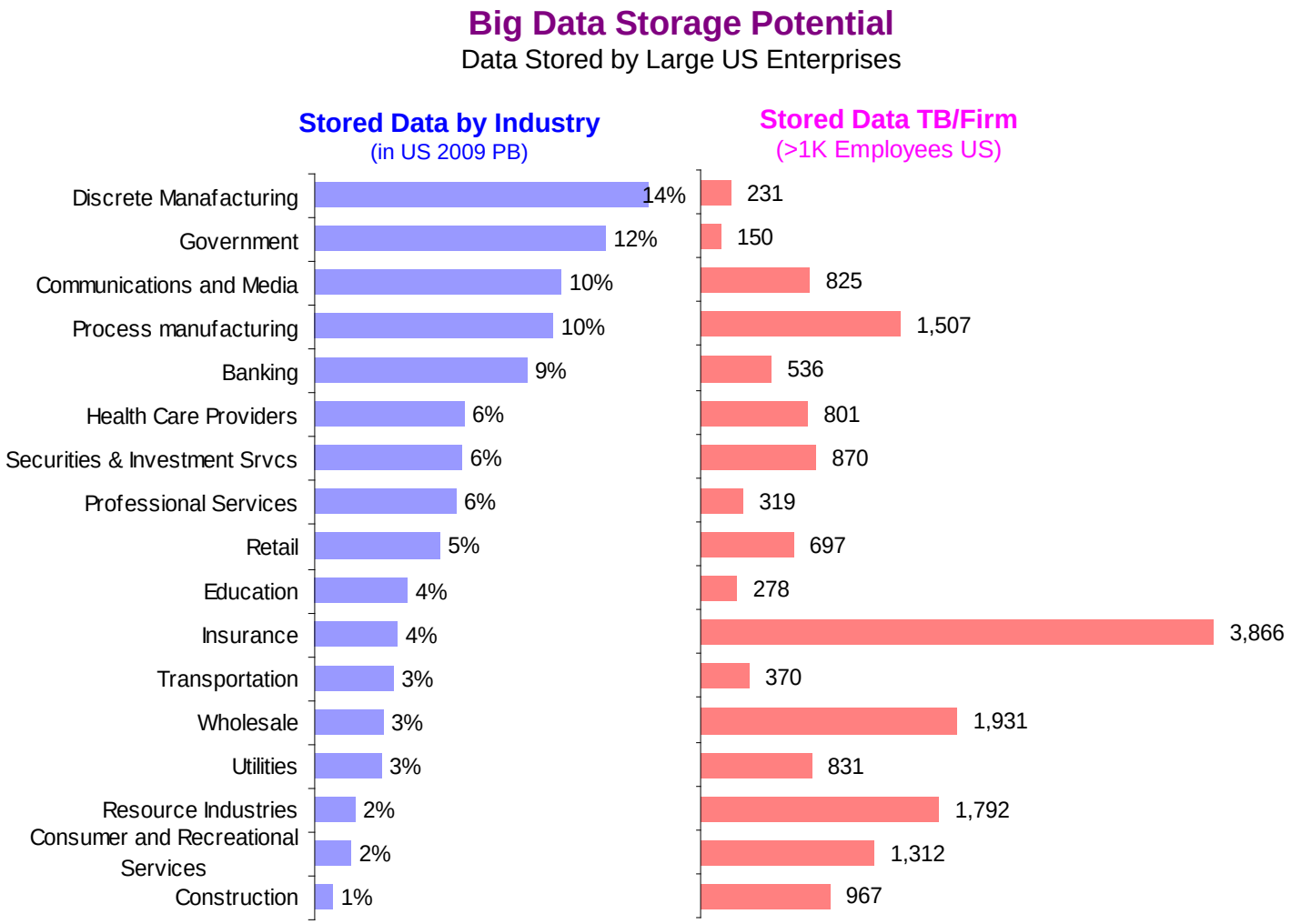
Value Potential of Using Big Data by Data Intensive Verticals



Big Data Targets: Storage Infrastructure

IMEX RESEARCH.COM

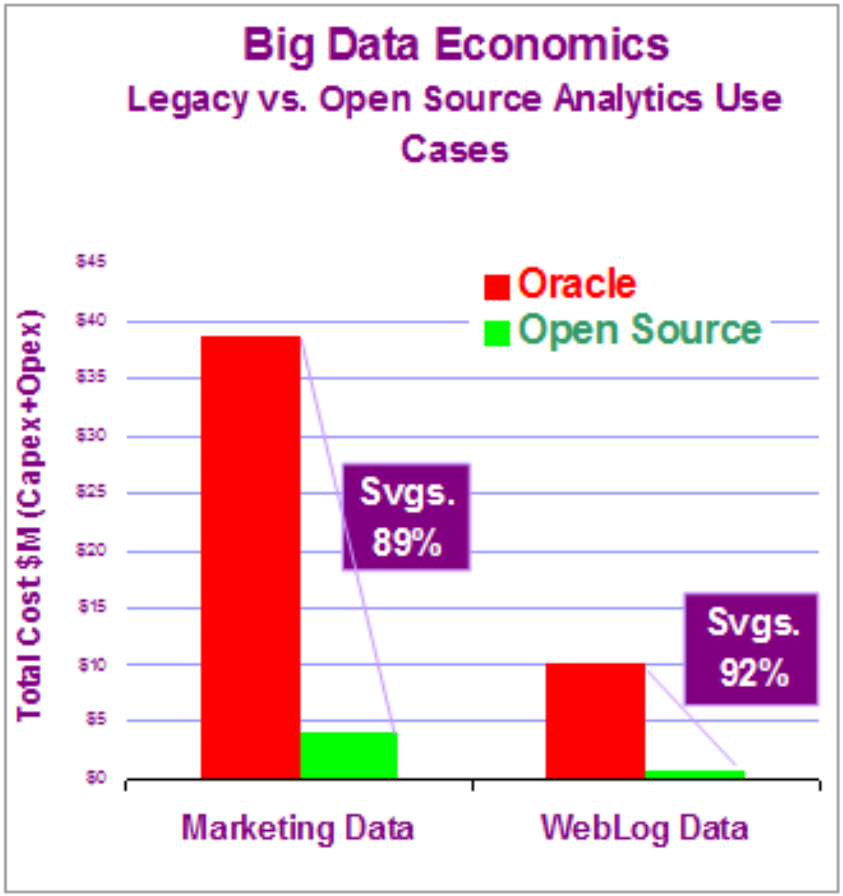
Data Stored by Large US Enterprises

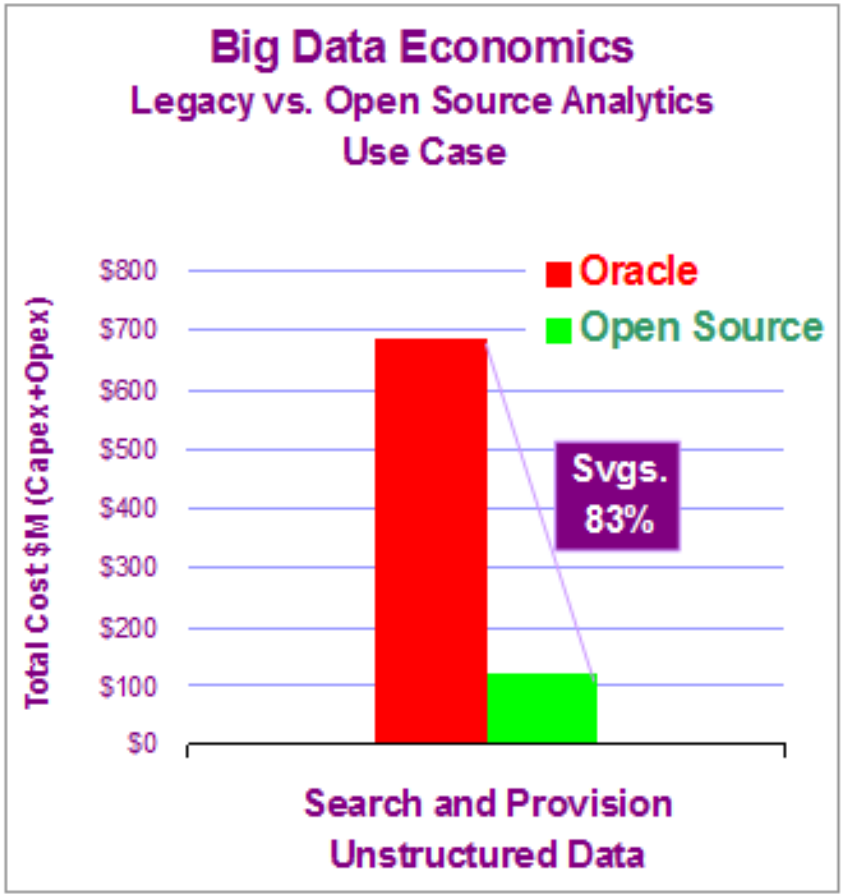


Big Data Targets: Savings w Open Source

IMEX RESEARCH.COM

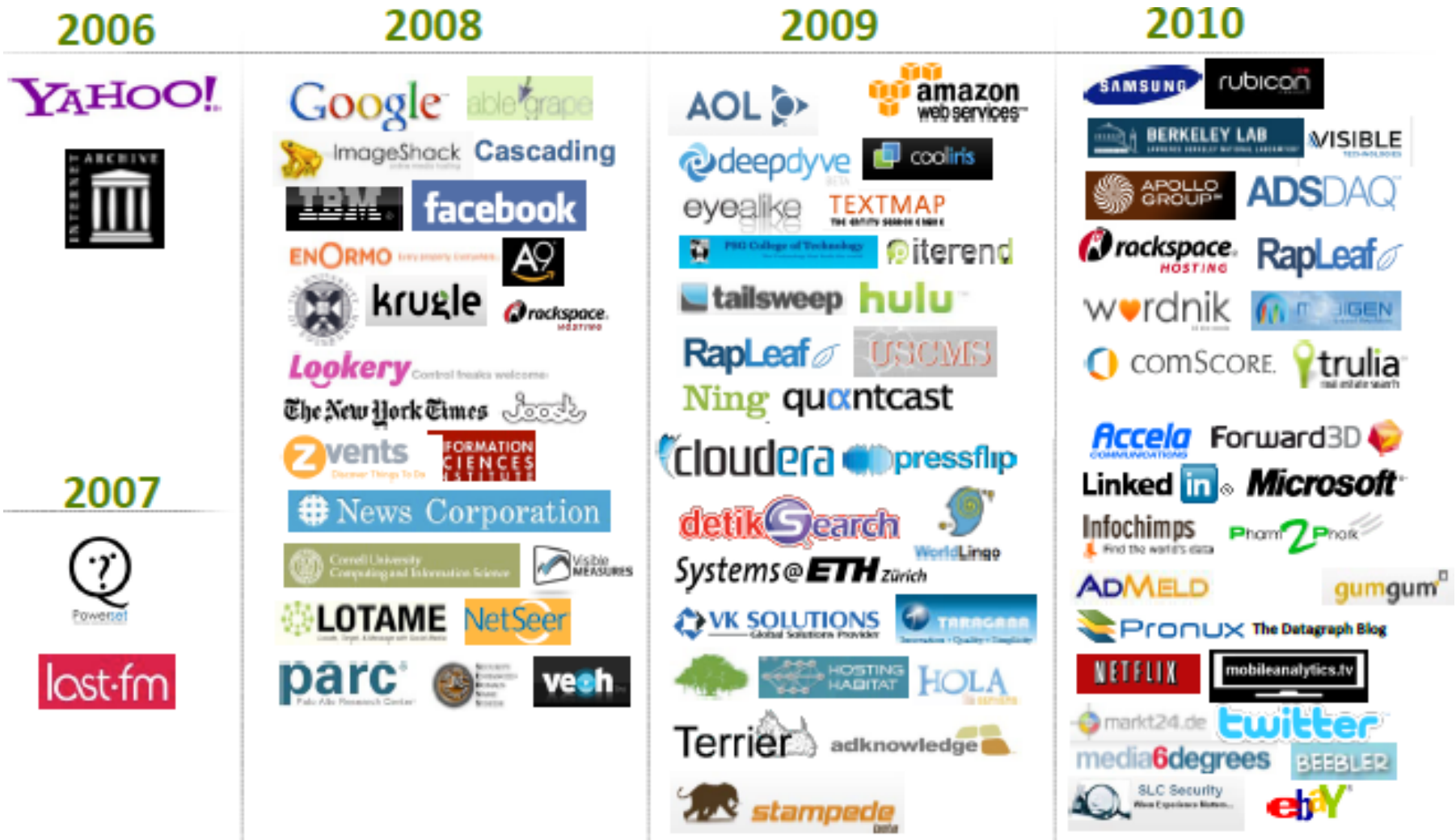
Legacy BI vs. Open Source Big Data Analytics





Rise of Big Data Adoption

IMEX RESEARCH.COM



Key Takeaways

IMEX RESEARCH.COM

- **Big Data creating paradigm shift in IT Industry**
  - Leverage the opportunity to optimize your computing infrastructure with Big Data Infrastructure after making a due diligence in selection of vendors/products, industry testing and interoperability.
  - Apply best storage technologies listed in this presentation and elsewhere
- **Optimize Big Data Analytics for Query Response Time vs. # of Users**
  - Improving Query Response time for a given number of users (IOPs) or Serving more users (IOPS) for a given query response time
- **Select Automated Storage Management Software**
  - **Data Forensics and Tiered Placement**
  - Every workload has unique I/O access signature
  - Historical performance data for a LUN can identify performance skews & hot data regions by LBAs.Non-disruptively migrate hot data using auto-tiering Software
- **Optimize Infrastructure to meet needs of Applications/SLA**
  - **Performance Economics/Benefits**
  - Typically 4-8% of data becomes a candidate and when migrated for higher performance tiering can provide response time reduction of ~65% at peak loads. Many industry Verticals and Applications will benefit using Big Data