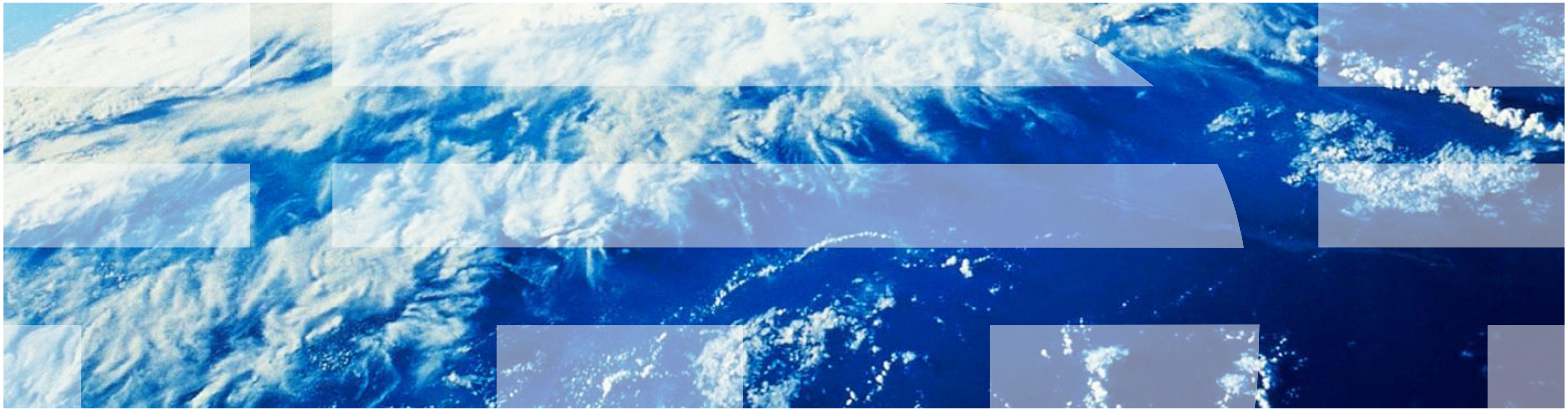
Big Data Concepts. Considerations

Gayane

IBM Client Center

82190210@ru.ibm.com

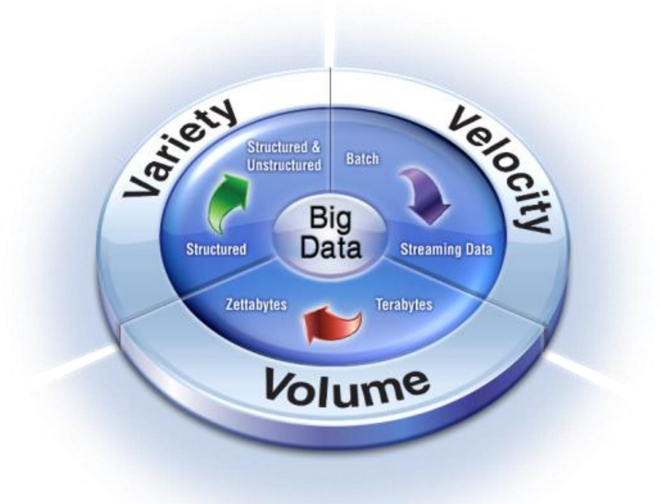


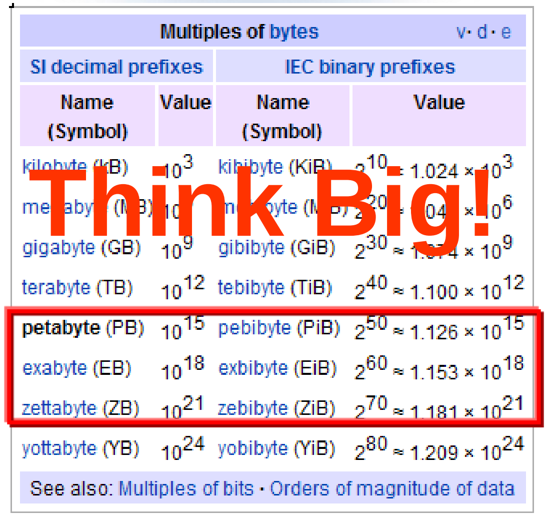
Agenda

- **What is Big Data?**
- **Data at Rest: Hadoop and InfoSphere BigInsights**
- **Data in Motion: InfoSphere Streams**
- **Considerations for BigInsights and Streams**
- **Concluding Thoughts**

The Big Data Opportunity

Extracting insight from an immense volume, variety and velocity of data, in a timely and cost-effective manner.





Multiples of bytes v · d · e			
SI decimal prefixes		IEC binary prefixes	
Name (Symbol)	Value	Name (Symbol)	Value
kilobyte (kB)	10^3	kibibyte (KiB)	$2^{10} \approx 1.024 \times 10^3$
megabyte (MB)	[illegible]	mebibyte (MiB)	$2^{20} \approx$ [illegible] $\times 10^6$
gigabyte (GB)	10^9	gibibyte (GiB)	$2^{30} \approx 1.074 \times 10^9$
terabyte (TB)	10^{12}	tebibyte (TiB)	$2^{40} \approx 1.100 \times 10^{12}$
petabyte (PB)	10^{15}	pebibyte (PiB)	$2^{50} \approx 1.126 \times 10^{15}$
exabyte (EB)	10^{18}	exbibyte (EiB)	$2^{60} \approx 1.153 \times 10^{18}$
zettabyte (ZB)	10^{21}	zebibyte (ZiB)	$2^{70} \approx 1.181 \times 10^{21}$
yottabyte (YB)	10^{24}	yobibyte (YiB)	$2^{80} \approx 1.209 \times 10^{24}$
See also: Multiples of bits · Orders of magnitude of data			

Think Big!

Variety: All kinds of data
All kinds of analytics

Velocity: Streaming data and large volume data movement

Volume: Scale from terabytes to zettabytes

Verasity Truthfulness a certainty of data

Where is this data coming from?

12+ TBs of tweet data every day

? TBs of data every day

25+ TBs of log data every day

30 billion RFID tags today (1.3B in 2005)

4.6 billion camera phones world wide

100s of millions of GPS enabled devices sold annually

76 million smart meters in 2009… 200M by 2014

2+ billion people on the Web by end 2011

What is “BIG DATA”? Where do I find it? Throw it away vs. Storing it?

Log files	practically every system creates and stores some kind of log	typically not examined unless there is trouble	Typically text and log wraps around on a regular basis to save space
HTTP	All web content	Web content based on xml format and is highly variable	No evident pattern to store in DB. DB cannot capture context
Metering and Instrumentation	Usually depicts real time status or cumulative value	Variable over time Usually collect over interval	Summary over interval or recorded peak. Patterns not analyzed
Video & Audio	Either streamed or stored in large files	Detail cannot be stored in DB	Only segments are of interest and require processing to analyze
Personal profiles	Volumes of texts and pictures	Can be external Can receive block or interval	Analysis requires cognitive – parsing Large volumes of retrieved data are irrelevant
Metadata	Information that is in addition to actual data stored in DB	Additional detail of a transaction that does not relate directly to billing	End to end story of events that relates to a transaction Large volumes of data that is difficult to store over time

*“**Big Data** technologies describe a new generation of technologies and architectures, designed to **economically extract value** from very large **volumes** of a wide **variety** of data, by enabling high **velocity** capture, discovery and/or analysis.”*

Source: Matt Eastwood, IDC

Concept Associations for Old Data and New Big Data

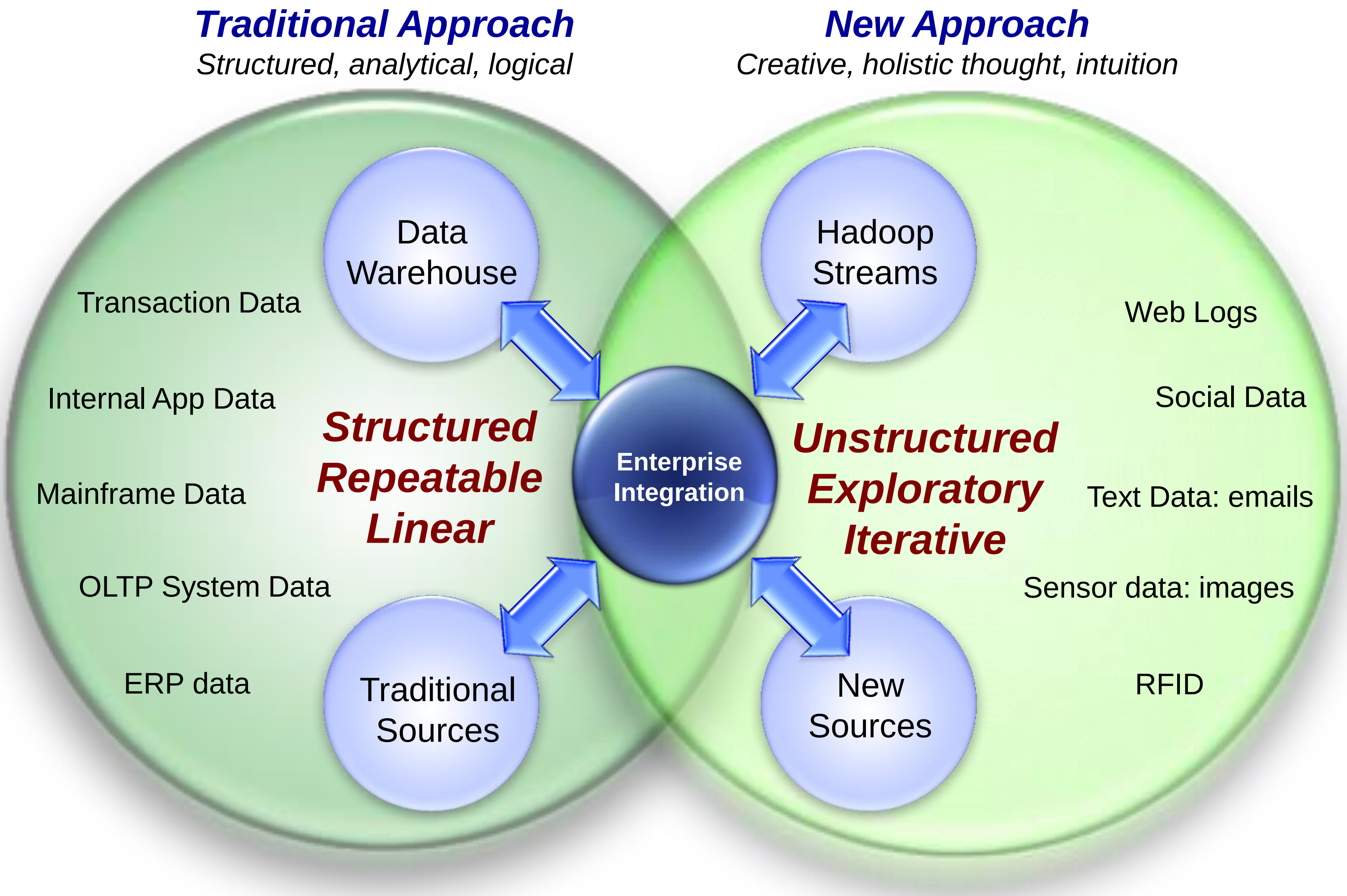
- Standard DBW (Warehouse)
  - Structured
  - Schema
  - Ad-hoc queries
  - Reports
  - Indexes
  - Repeatable
  - Optimized queries
  - ETL
  - Cleansed data
  - Transactions
  - High availability
  - MPP
  - SMP
  - Complex analytics
  - Data models
  - Master data
  - Model building
  - SQL

- New Big Data (BigInsights, Streams)
  - Unstructured
  - Streaming
  - Discovery
  - Programming
  - Text analytics
  - Video
  - Time series
  - Sensors
  - Log files
  - Noisy data
  - Commodity hardware
  - Cluster
  - Real-time analytics
  - Complex analytics
  - Tweets
  - Sentiment analysis
  - Social network analysis
  - Model-driven optimization
  - NoSQL

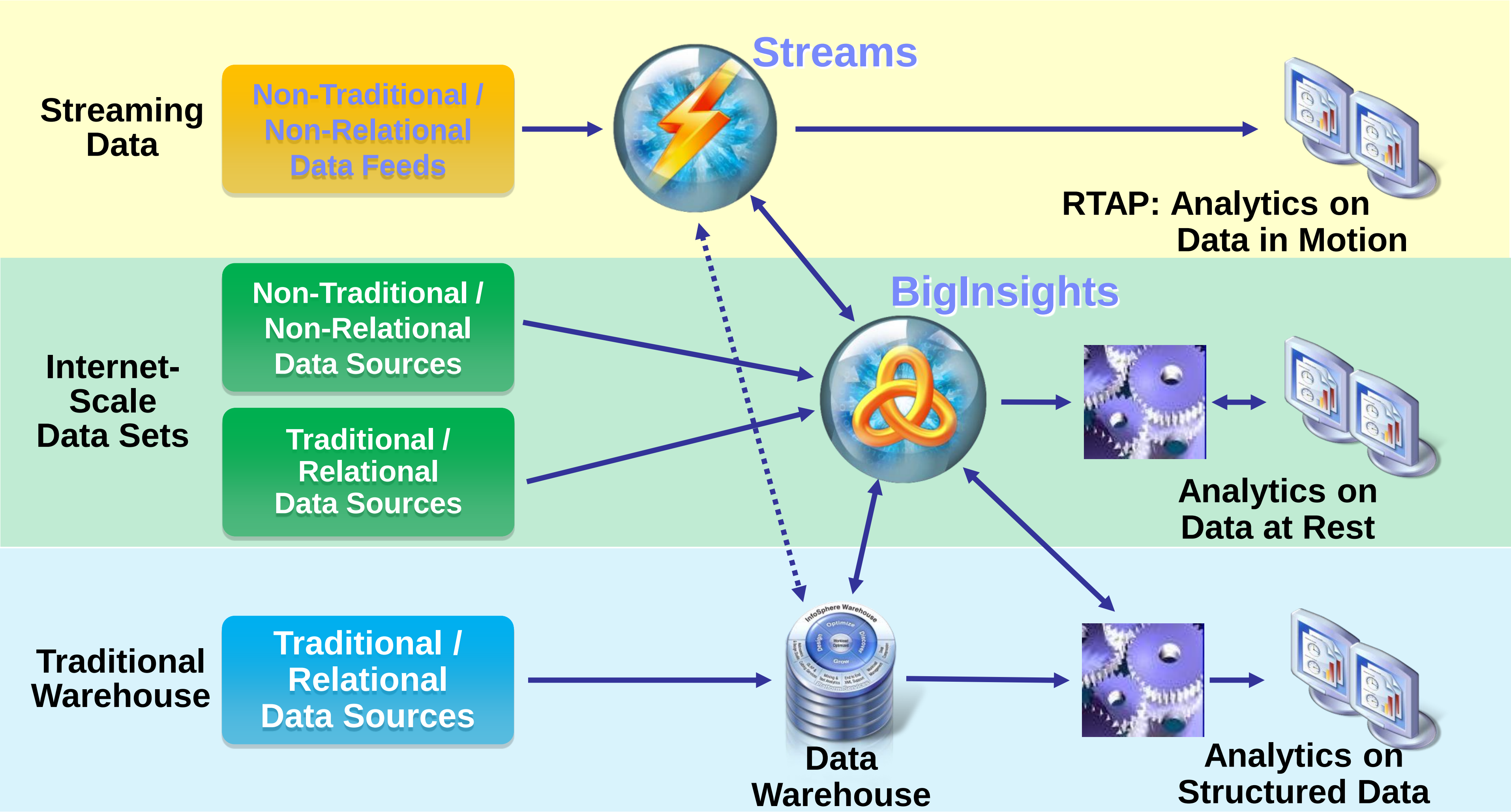
Data Warehouse and BigInsights Comparison Chart

	Data Warehouse	Hadoop / Streams
Data Types	Largely structured data	Any type of data, structured or unstructured
Data Loading	Data is cleansed/structured before going into the warehouse to maximize its utility	Raw data may be ingested as is, without any modification, as the relationships may not be understood or defined
Reliability	ACID compliant	Not ACID compliant
Integrity	Database maintains integrity	Applications code integrity
Analytic Approach	▪ *High value*, structured data ▪ *Repeated* operations and processes (e.g. transactions, reports, BI, etc.) ▪ Relatively *stable* sources ▪ Well-understood requirements ▪ Optimized for fast access and analysis	▪ *Highly variable* data and content ▪ *Iterative*, exploratory analysis (e.g. scientific research, behavioral modeling) ▪ *Volatile* sources ▪ Ill-defined questions and changing requirements ▪ Optimized for flexibility
Hardware	Powerful appliance and optimized hardware	Inexpensive, commodity hardware

IBM’s Value: Complementary Analytics



The Big Data Ecosystem: Interoperability is Key



Streaming Data

Non-Traditional / Non-Relational Data Feeds

Streams

RTAP: Analytics on Data in Motion

Internet-Scale Data Sets

Non-Traditional / Non-Relational Data Sources

Traditional / Relational Data Sources

BigInsights

Analytics on Data at Rest

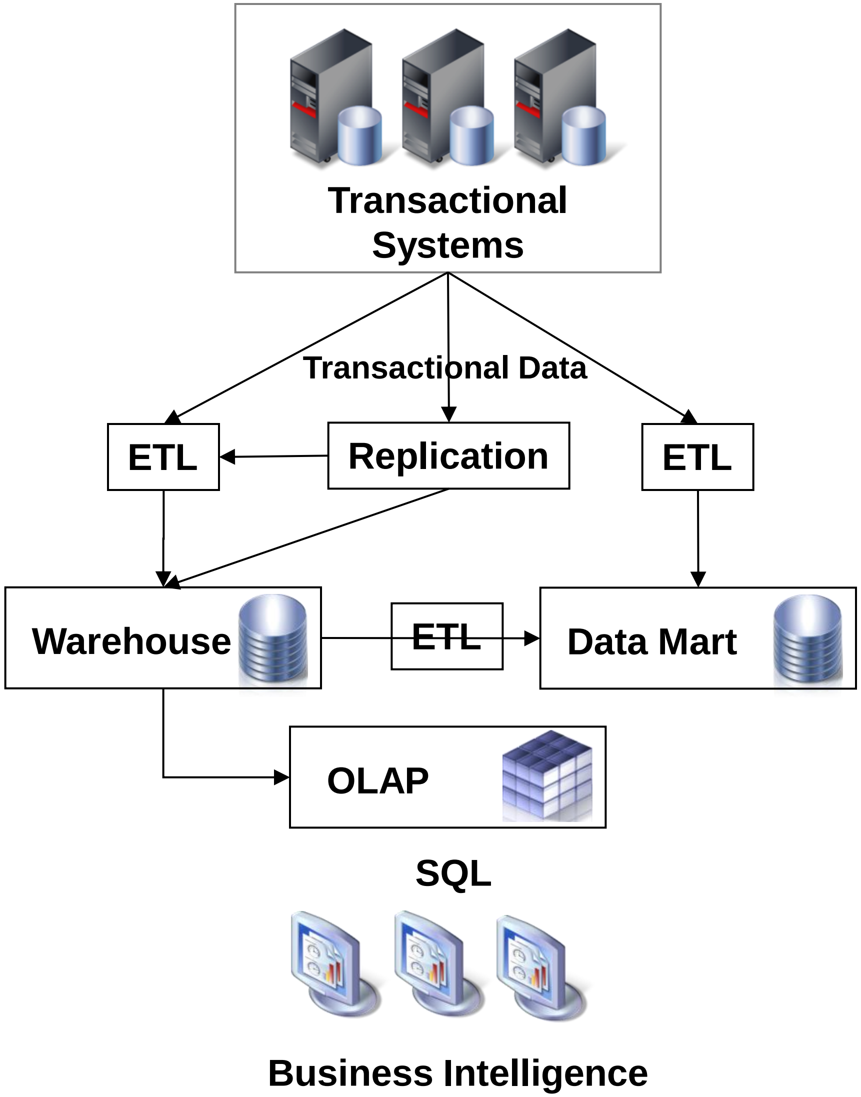
Traditional Warehouse

Traditional / Relational Data Sources

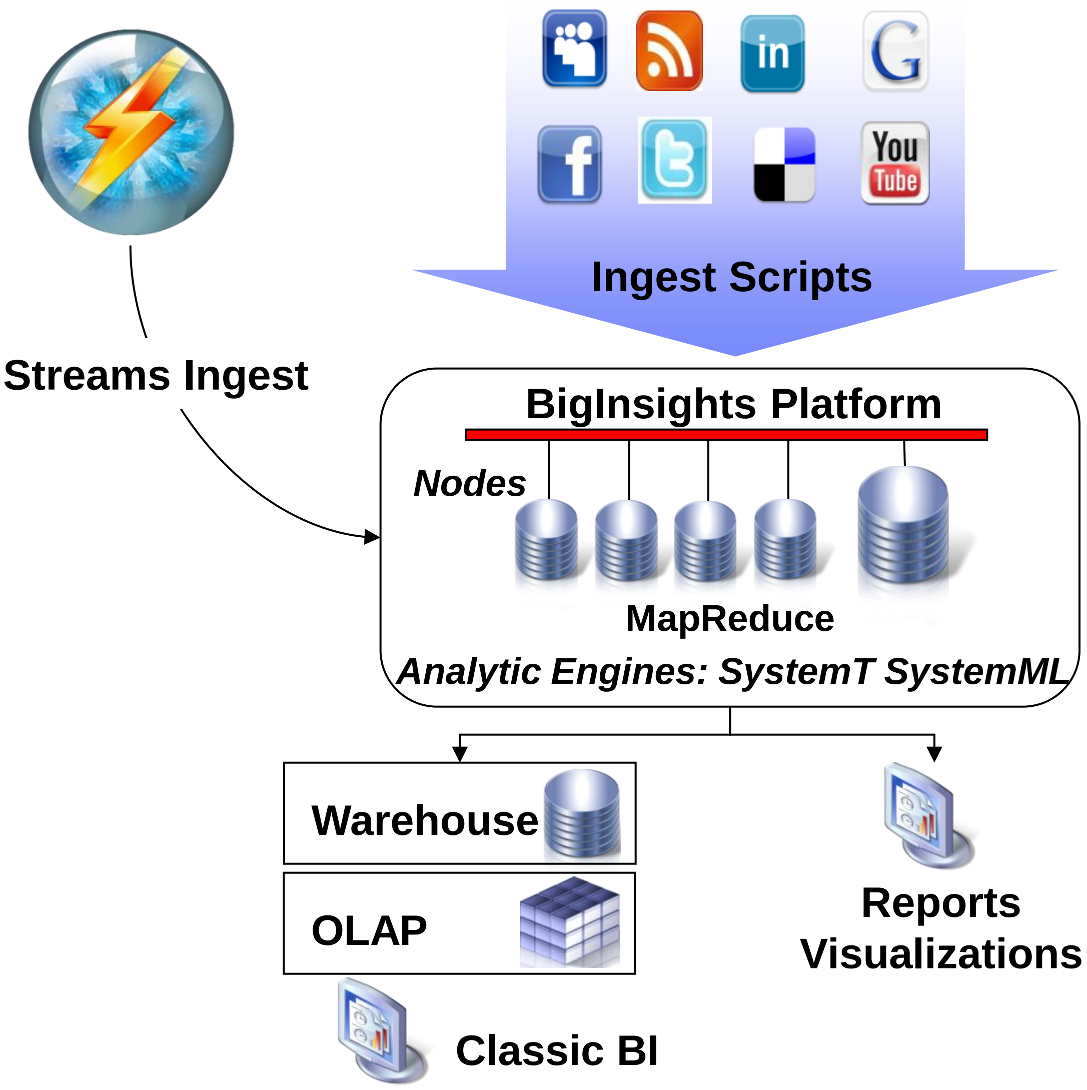
Data Warehouse

Analytics on Structured Data

Classic OLTP/Data Warehouse Environment



Hadoop / Streams Environment


Streams Ingest

Ingest Scripts

BigInsights Platform

Nodes

MapReduce

Analytic Engines: SystemT SystemML

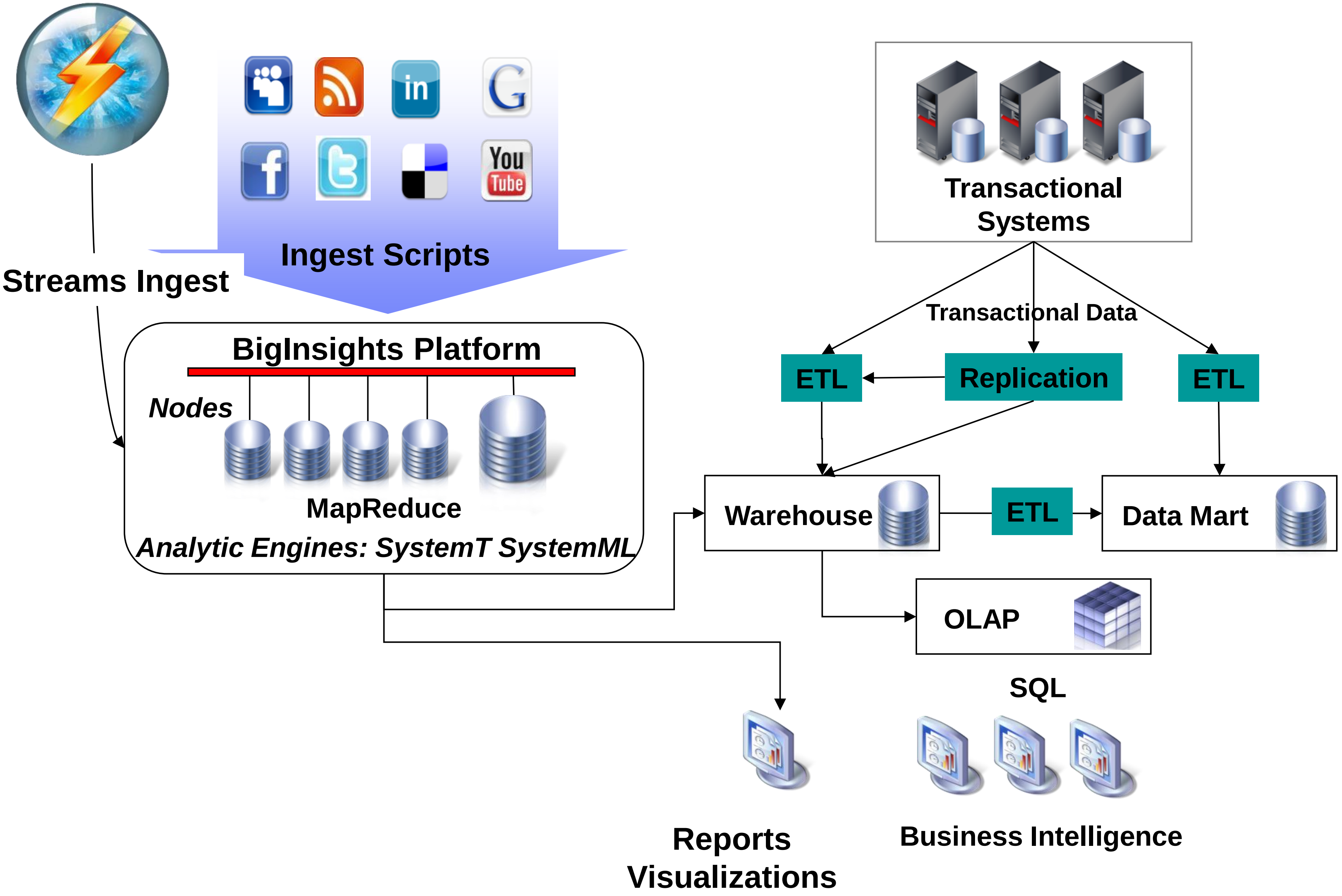
Warehouse

OLAP

Classic BI

Reports
Visualizations

New Consolidated Environment



What can you do with big data?

Financial Services

- Fraud detection
- Risk management
- 360° View of the Customer

Utilities

- Weather impact analysis on power generation
- Transmission monitoring
- Smart grid management

Transportation

- Weather and traffic impact on logistics and fuel consumption
- Traffic congestion

IT

- System log analysis
- Cybersecurity

Health & Life Sciences

- Epidemic early warning
- ICU monitoring
- Remote healthcare monitoring

Retail

- 360° View of the Customer
- Click-stream analysis
- Real-time promotions

Telecommunications

- CDR processing
- Churn prediction
- Geomapping / marketing
- Network monitoring

Law Enforcement

- Real-time multimodal surveillance
- Situational awareness
- Cyber security detection

The IBM Big Data Platform

Apache Hadoop: open source framework for the distributed processing of large data sets across clusters of computers using a simple programming model

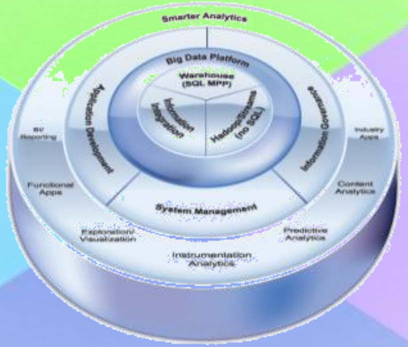
InfoSphere BigInsights

Hadoop-based low latency analytics for variety and volume

Data-At-Rest

Hadoop

Information Integration



Stream Computing

InfoSphere Information Server

High volume data integration and transformation

InfoSphere Streams

Low Latency Analytics for streaming data

Velocity, Variety & Volume

Data-In-Motion

MPP Data Warehouse

Netezza High Capacity Appliance

PureData for Analytics

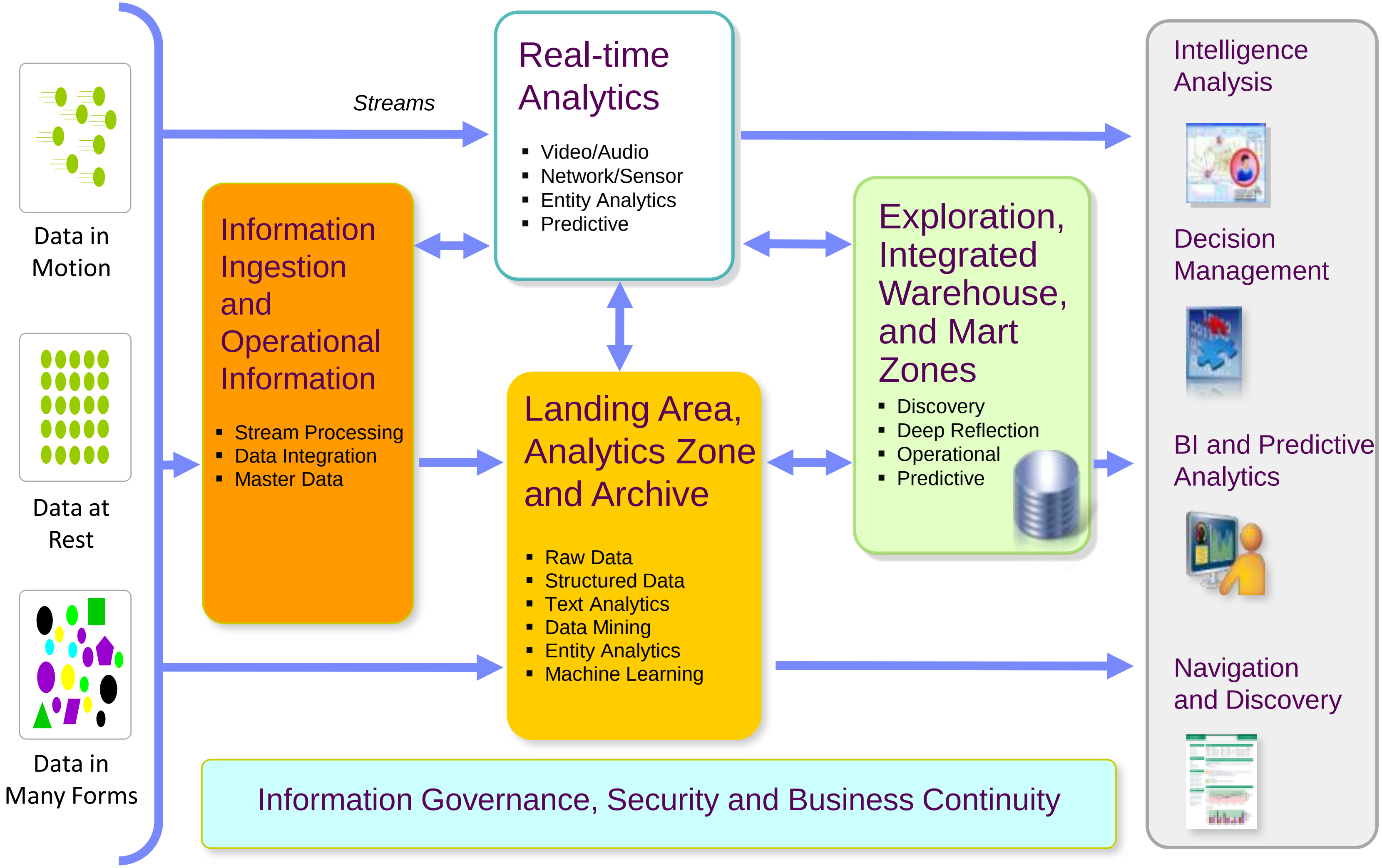
InfoSphere Warehouse

Large volume structured data analytics

Informix Timeseries

Time-structured analytics

New Architecture to Leverage All Data and Analytics



Data in Motion

Data at Rest

Data in Many Forms

Streams

Real-time Analytics

- Video/Audio
- Network/Sensor
- Entity Analytics
- Predictive

Information Ingestion and Operational Information

- Stream Processing
- Data Integration
- Master Data

Landing Area, Analytics Zone and Archive

- Raw Data
- Structured Data
- Text Analytics
- Data Mining
- Entity Analytics
- Machine Learning

Exploration, Integrated Warehouse, and Mart Zones

- Discovery
- Deep Reflection
- Operational
- Predictive

Intelligence Analysis

Decision Management

BI and Predictive Analytics

Navigation and Discovery

Information Governance, Security and Business Continuity

Entry points are accelerated by products within the big data platform

1 – Discover, understand and navigate all big data sources

Data Explorer (Vivisimo)

2 – Take advantage of streaming data and find the valuable insights

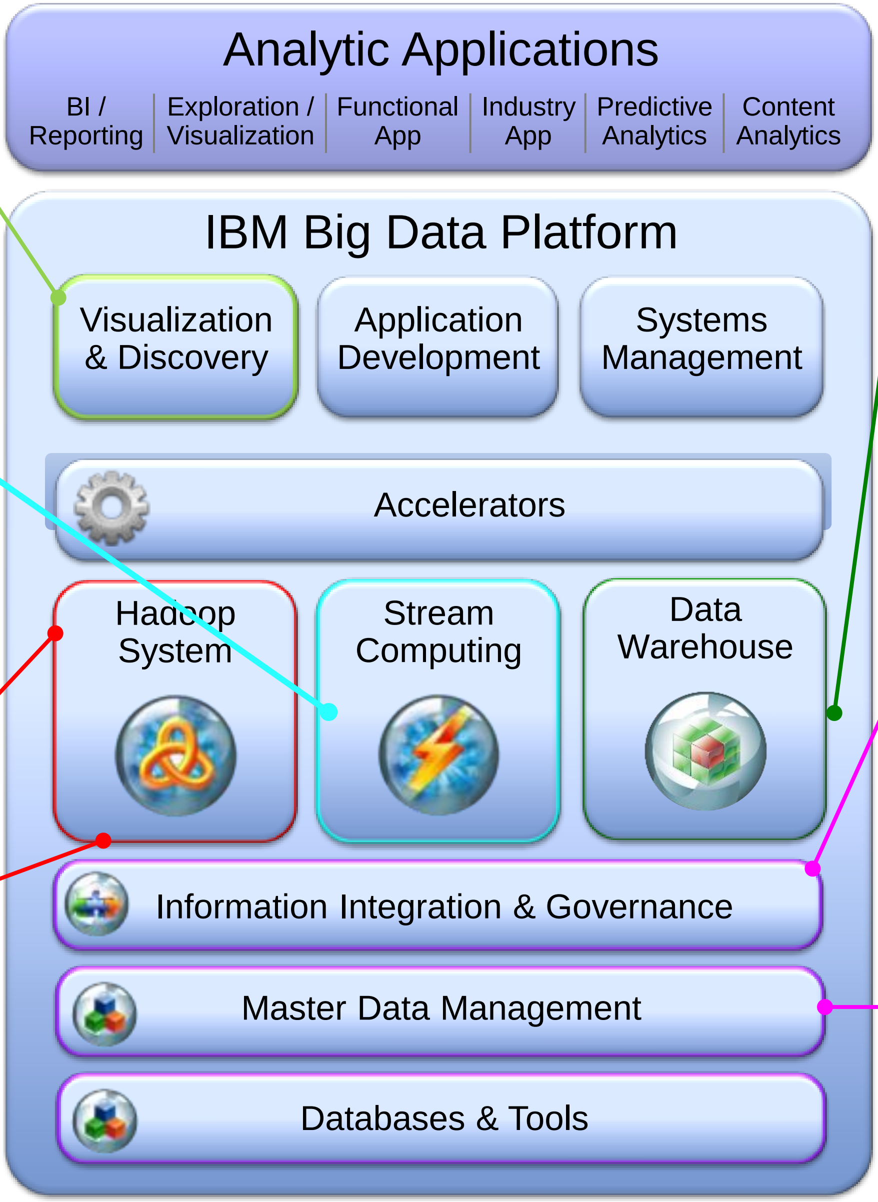
InfoSphere Streams

3 - Extend your data warehouse to incorporate new data types

InfoSphere BigInsights

4 – Reduce the cost of your data warehouse by utilizing Hadoop

InfoSphere BigInsights



Analytic Applications

BI / Reporting | Exploration / Visualization | Functional App | Industry App | Predictive Analytics | Content Analytics

IBM Big Data Platform

Visualization & Discovery

Application Development

Systems Management

Accelerators

Hadoop System

Stream Computing

Data Warehouse

Information Integration & Governance

Master Data Management

Databases & Tools

5 – Optimize infrastructure to lower cost and improve performance

PDA/PDOA (Netezza/ISAS)

6 – Establish and maintain an accurate, secure and trusted view of all available data

InfoSphere Information Server & Optim

7 – Manage and leverage trusted customer and product information

InfoSphere MDM

Big Data - Hadoop

What is Hadoop?

Traditonal computation model

- Bring data to the function
- Load data into memory, and process on a central server
- Does not scale well for Big Data problems

Apache Hadoop: open source framework for data-intensive applications

- Inspired by Google technologies (MapReduce, GFS)
- Well-suited to batch-oriented, read-intensive applications
- Yahoo! Adopted these technologies and open sourced them into the Apache Hadoop project

Enables applications to work with thousands of nodes and petabytes of data in a highly parallel, cost effective manner

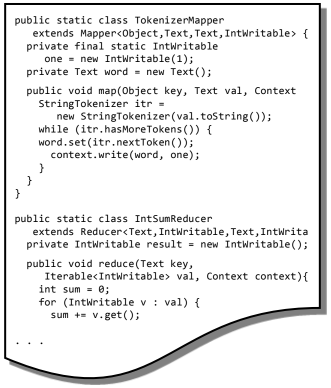
- CPU + disks of commodity box = Hadoop “node”
- Boxes can be combined into massive clusters
- New nodes can be added as needed without changing
  - Data formats
  - How data is loaded
  - How jobs are written

Hadoop Explained, Two Key Concepts: Map Reduce

Hadoop computation model

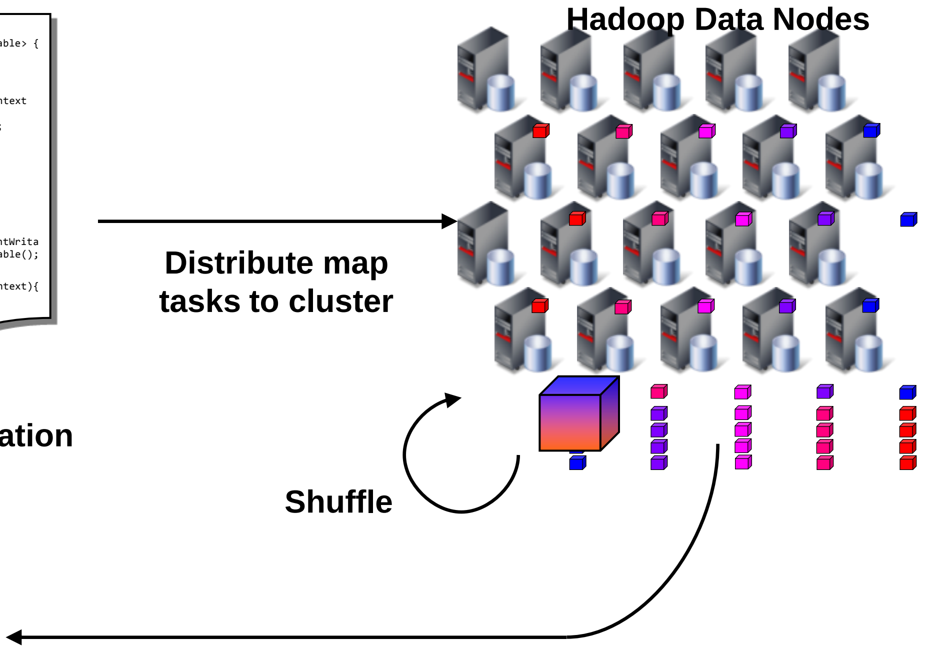
- Data stored in a distributed file system spanning many inexpensive computers
- Bring function to the data
- Distribute application to the compute resources where the data is stored

Scalable to thousands of nodes and petabytes of data



```
public static class TokenizerMapper
   extends Mapper<Object,Text,Text,IntWritable> {
  private final static IntWritable
     one = new IntWritable(1);
  private Text word = new Text();

  public void map(Object key, Text val, Context
    StringTokenizer itr =
       new StringTokenizer(val.toString());
    while (itr.hasMoreTokens()) {
    word.set(itr.nextToken());
      context.write(word, one);
    }
  }
}

public static class IntSumReducer
   extends Reducer<Text,IntWritable,Text,IntWrita
  private IntWritable result = new IntWritable();

  public void reduce(Text key,
     Iterable<IntWritable> val, Context context){
    int sum = 0;
    for (IntWritable v : val) {
      sum += v.get();

. . .
```

MapReduce Application



Hadoop Data Nodes

Distribute map tasks to cluster

Shuffle

Result Set

Return a single result set

1. Map Phase
   (break job into small parts)
2. Shuffle
   (transfer interim output for final processing)
3. Reduce Phase
   (boil all output down to a single result set)

Hadoop Explained, Two Key Concepts: HDFS

HDFS stores data across multiple nodes

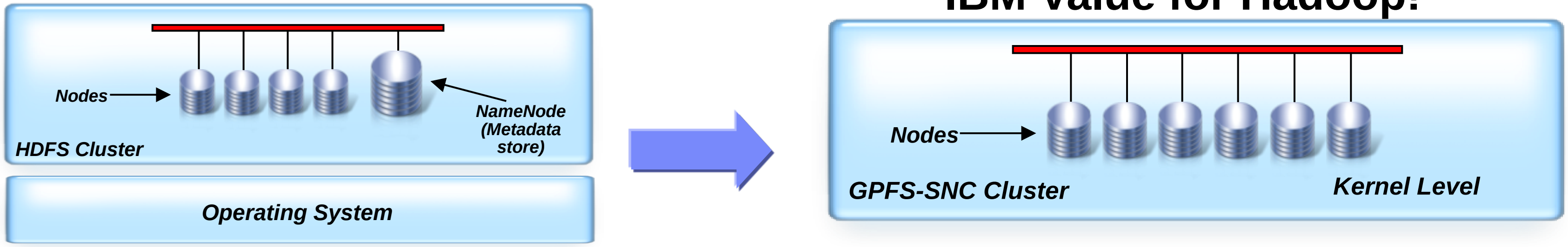
HDFS achieves reliability by replicating data across multiple nodes (typically 3 or more)

file system is a cluster of data nodes

serves up blocks of data over the network using a block protocol specific to HDFS

HDFS Name Node is a single point of failure

IBM eliminates this single point of failure while improving file system performance with GPFS-SNC (available as beta code)



IBM Value for Hadoop!

InfoSphere BigInsights = Hadoop + IBM Innovation

hadoop + **IBM Innovation**

Scalable

- New nodes can be added on the fly

Affordable

- Massively parallel computing on commodity servers

Flexible

- Hadoop is schema-less, and can absorb any type of data

Fault Tolerant

- Through MapReduce software framework

Performance & Reliability

- Adaptive MapReduce, Compression, BigIndex, Flexible Scheduler

Analytic Accelerators

Productivity Accelerators

- Web-based UIs
- Tools to leverage existing skills
- End-user visualization

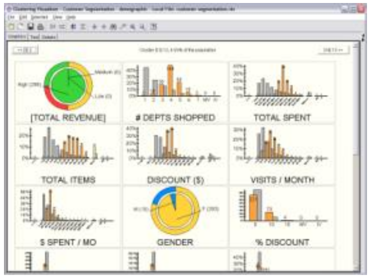
Enterprise Integration

- To extend & enrich your information supply chain

How IBM BigInsights extends Hadoop capability

Manageability

- Single Click Integrated Install
- Browser Based Cluster Mgmt
- GPFS-SNC



Developer Value

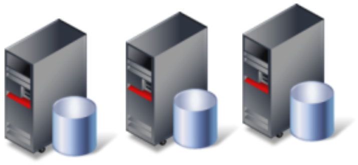
- New Query Language (jaql)
- Eclipse Tools for Analytics
- Broad Integration with other Information Management Technologies (DW, DataStage, RDBMS, et al)
- Integration with InfoSphere Streams

Advanced Analytics

- Bundled Scalable Text Analytics (AQL – Sentiment Analysis – NLP)
- BigData Scale Visualization (BigSheets)
- Bundled Scalable Machine Learning (DML)

Performance & Availability

- GPFS-SNC (Data Replication)
- Splittable Compression
- Improved Map/Reduce
- Job Scheduling Improvements
- Large Scale Indexing



Security

- Secure File System (GPFS-SNC)
- Cluster Hardening

Providing competitive advantage

Faster analysis, design and simulation while managing costs

Financial Services

Banking, financial markets, insurance

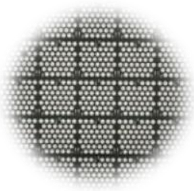
Risk management, compliance, investment decisions, liquidity management

Manufacturing

Automotive, Aerospace and Defense and Engineering

New designs, more complex products and higher quality

Electronics



Electronics and Semiconductor

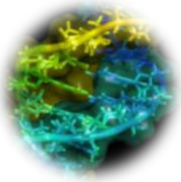
More complex designs and simulations

Petroleum

Oil and gas exploration and production

Reserve identification, imaging and reclamation

Research & Academic



Life sciences, research, higher education

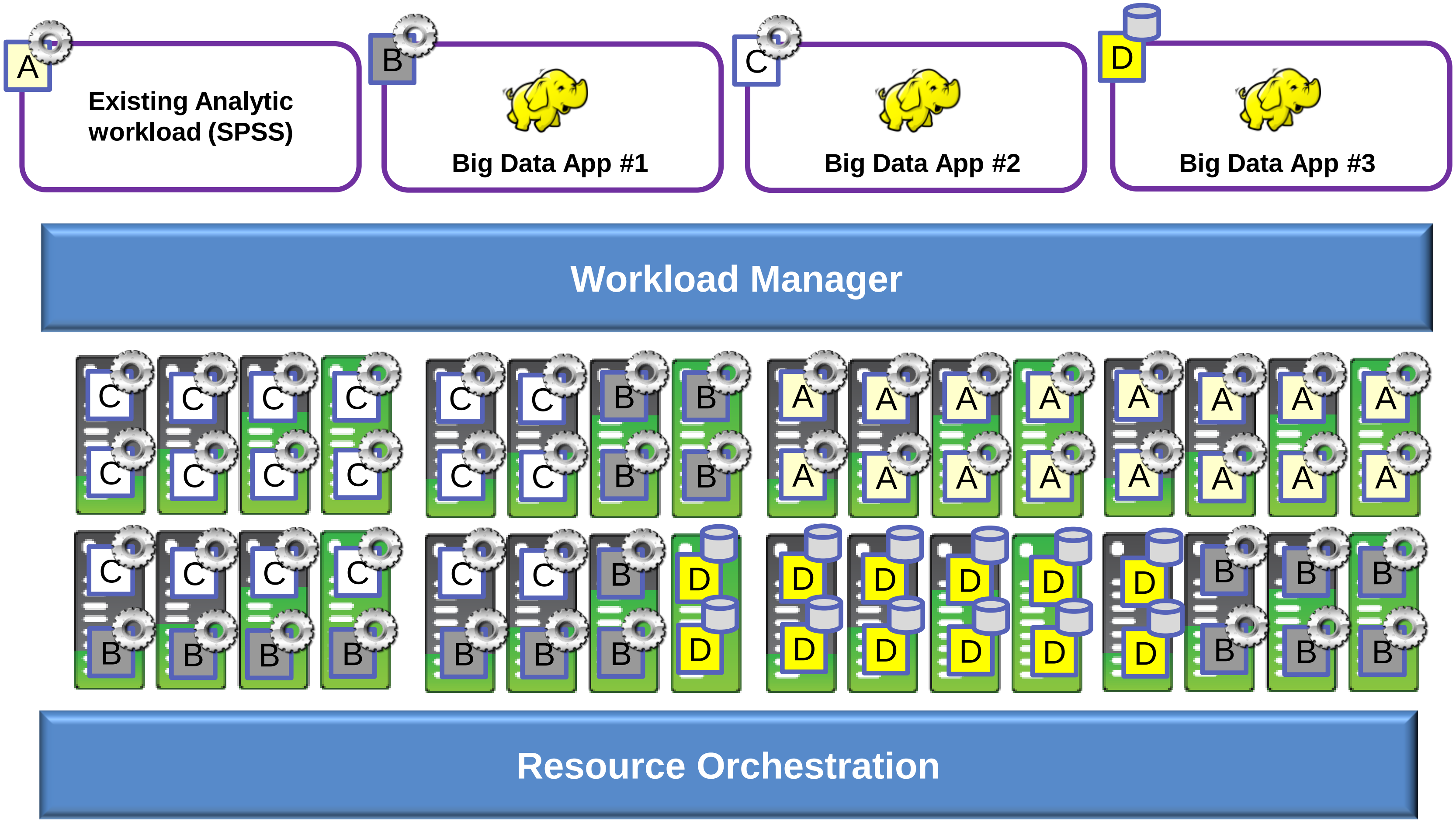
Drug development, sequencing, cross department collaboration

Government & Intelligence

Scientific research, classified/defense, weather/ environmental sciences

Intelligence gathering and insight development

Dynamic resource sharing among heterogeneous tenants



Existing Analytic workload (SPSS)

Big Data App #1

Big Data App #2

Big Data App #3

Workload Manager

Resource Orchestration

A
B
C
D
C C C C C C B B A A A A A A A A
C C C C C C B B A A A A A A A A
C C C C C C B D D D D D D B B B
B B B B B B B D D D D D D B B B

Streams

Stream Computing – Analyze Data in Motion

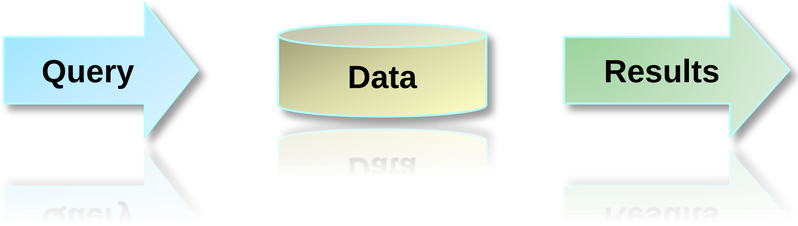
Traditional Computing

Historical fact finding

Find and analyze information stored on disk

Batch paradigm, pull model

Query-driven: submits queries to static data



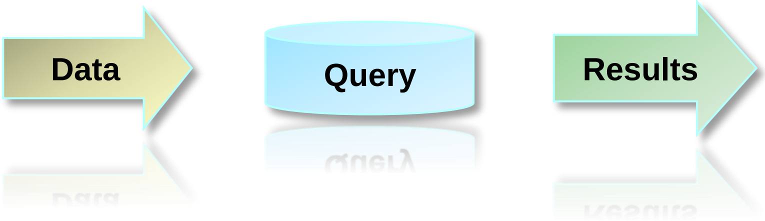
Stream Computing

Current fact finding

Analyze data in motion – before it is stored

Low latency paradigm, push model

Data driven: bring the data to the query



InfoSphere Streams: Massively Scalable Stream Analytics

Linear Scalability

- Clustered deployments – unlimited scalability

Automated Deployment

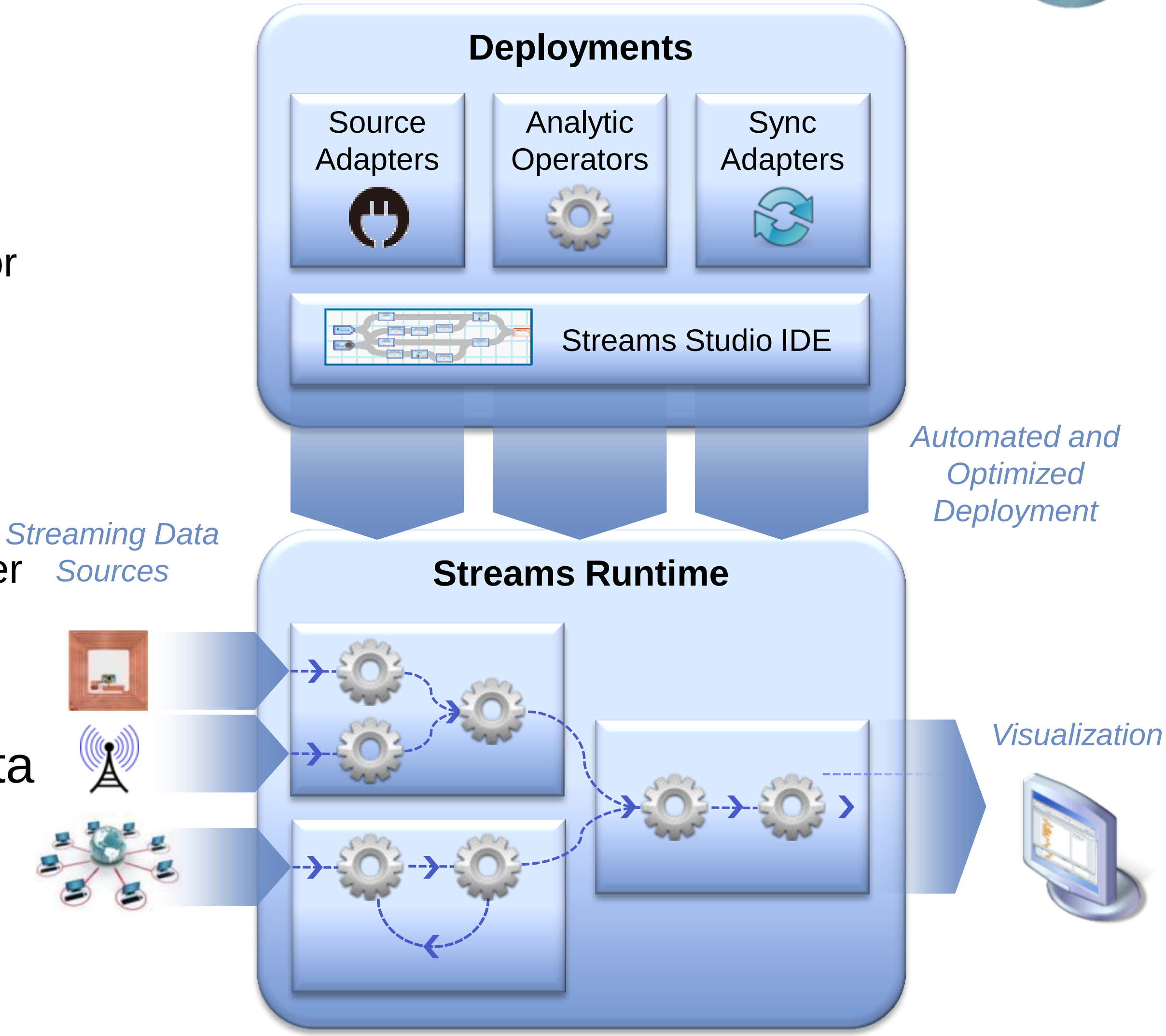
- Automatically optimize operator deployment across clusters

Performance Optimization

- JVM Sharing – minimize memory use
- Fuse operators on same cluster
- Telco client – 25 Million messages per second

Analytics on Streaming Data

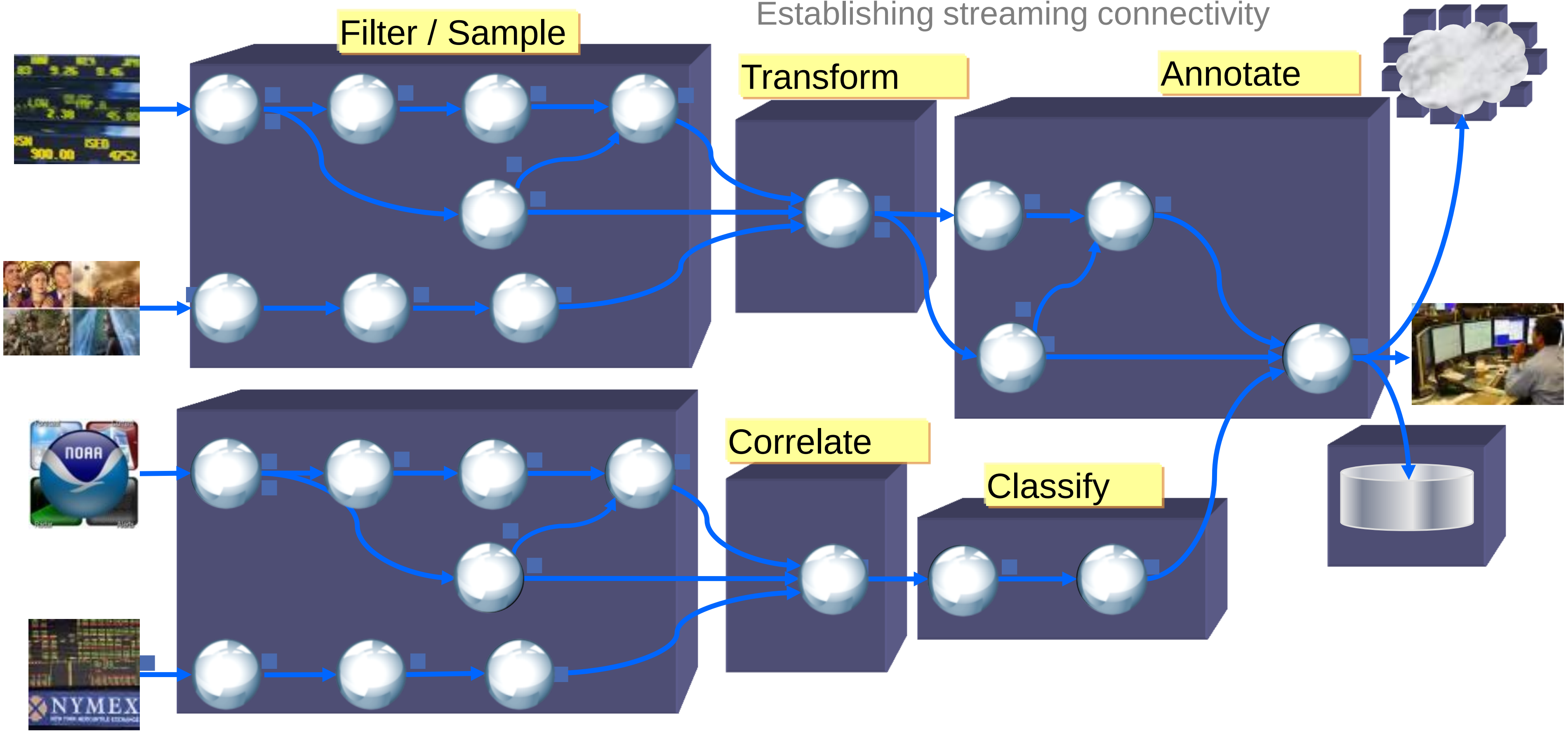
- Analytic accelerators for a variety of data types (text, acoustic, image, video, geospatial, etc)
- Optimized for real-time performance



How Streams Works

→ Continuous ingestion
→ Continuous analysis

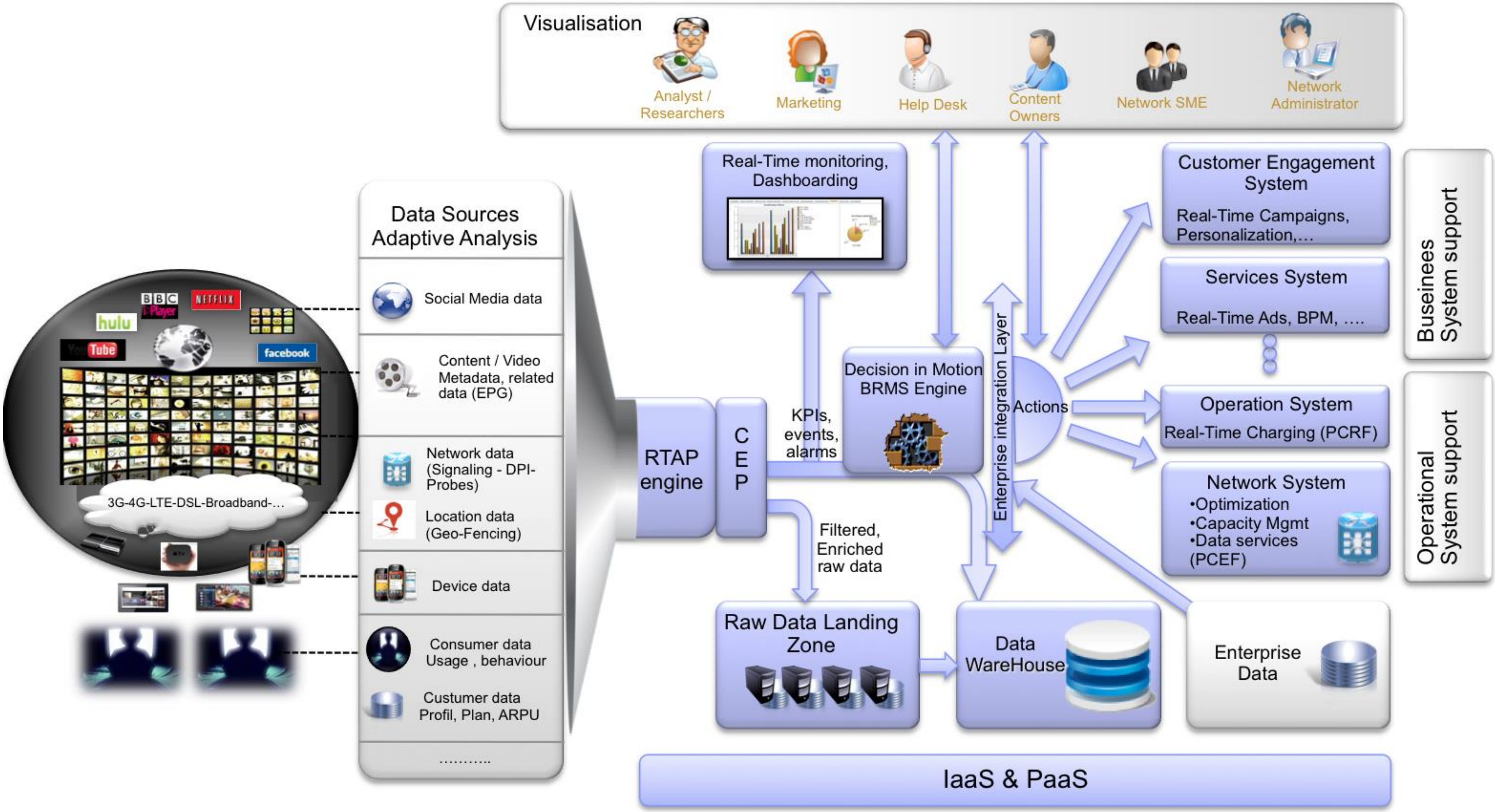
Infrastructure provides services for
Scheduling analytics across hardware hosts,
Establishing streaming connectivity



Achieve scale:
By partitioning applications into software components
By distributing across stream-connected hardware hosts

Where appropriate:
Elements can be *fused* together
for lower communication latency

IBM Consumer Oriented Analytic Architecture


Visualisation
Analyst / Researchers
Marketing
Help Desk
Content Owners
Network SME
Network Administrator
Real-Time monitoring, Dashboarding
Customer Engagement System
Real-Time Campaigns, Personalization,…
Services System
Real-Time Ads, BPM, ….
Buseinees System support
Operation System
Real-Time Charging (PCRF)
Network System
•Optimization
•Capacity Mgmt
•Data services (PCEF)
Operational System support
BBC iPlayer
NETFLIX
hulu
YouTube
facebook
3G-4G-LTE-DSL-Broadband-…
Data Sources Adaptive Analysis
Social Media data
Content / Video Metadata, related data (EPG)
Network data (Signaling - DPI-Probes)
Location data (Geo-Fencing)
Device data
Consumer data Usage , behaviour
Custumer data Profil, Plan, ARPU
………..
RTAP engine
C E P
KPIs, events, alarms
Decision in Motion BRMS Engine
Enterprise integration Layer
Actions
Filtered, Enriched raw data
Raw Data Landing Zone
Data WareHouse
Enterprise Data
IaaS & PaaS

CEP : Complex Event Processing RTAP : Real-Time Analytic Processing
BRMS : Business Rules Management System

Industrial references

Banking

- Customer Intimacy & Offer Optimization
- Integration into Enterprise Risk

Insurance

- Cross-sell & Retention
- Risk-based Claims Processing

Government

- Crime Prevention
- Analytics for Smarter Cities

Education

- Student Performance

Retail

- Market Basket Analysis
- Assortment Planning

Telco

- Churn Management

Industrial

- Predictive Maintenance

Thank You

Your questions