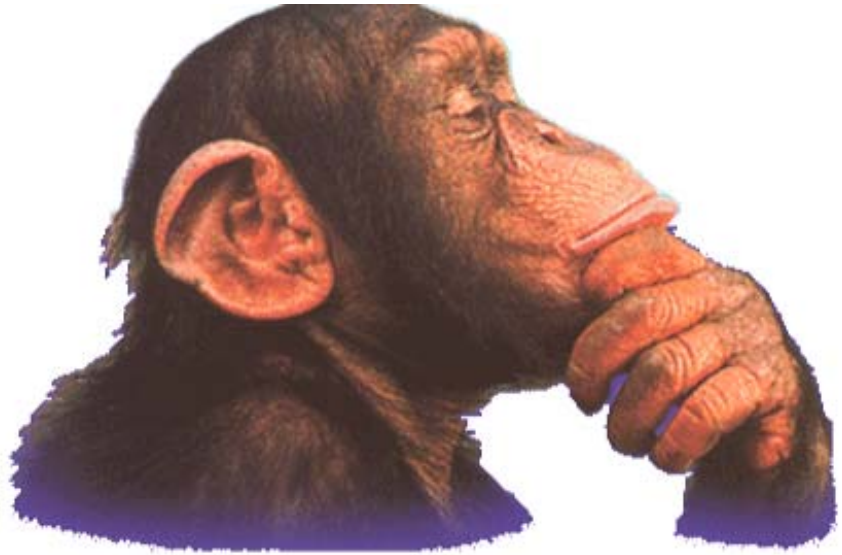
SoftServe

Big Data Analytics

Reference Architectures and Case Studies

Relational vs. Non-Relational Architecture

Relational

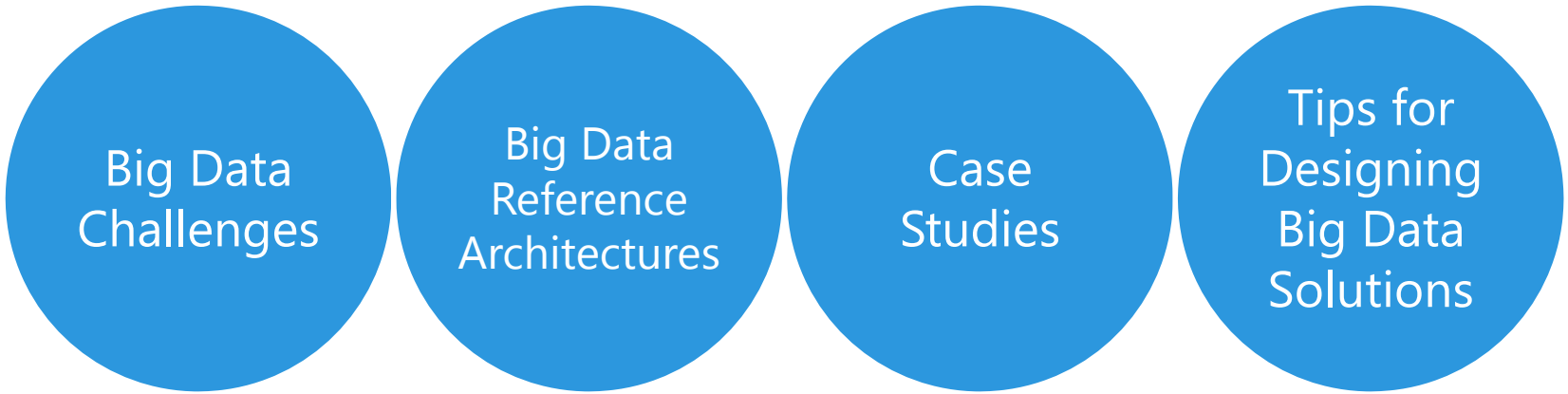


- Rational
- Predictable
- Traditional

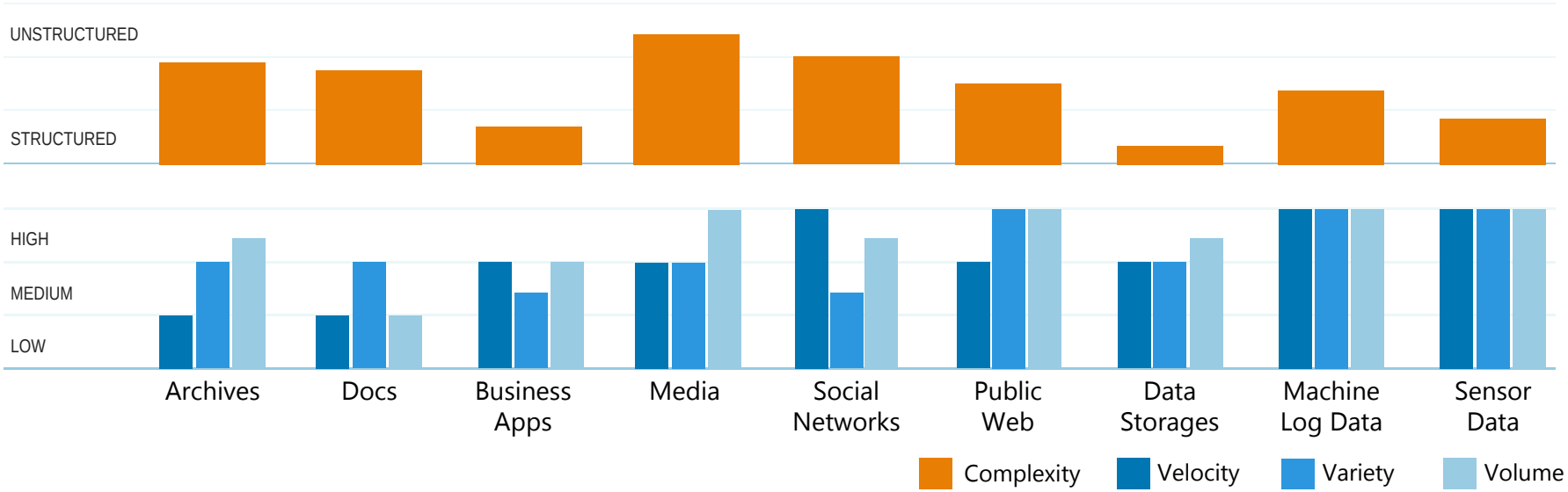
Non-Relational

- Agile
- Flexible
- Modern

Agenda


Big Data Challenges
Big Data Reference Architectures
Case Studies
Tips for Designing Big Data Solutions

Big Data Challenges



Archives
Scanned documents, statements, medical records, e-mails etc..

Docs
XLS, PDF, CSV, HTML, JSON etc.

Business Apps
CRM, ERP systems, HR, project management etc.

Media
Images, video, audio etc.

Social Networks
Twitter, Facebook, Google+, LinkedIn etc.

Public Web
Wikipedia, news, weather, public finance etc

Data Storages
RDBMS, NoSQL, Hadoop, file systems etc.

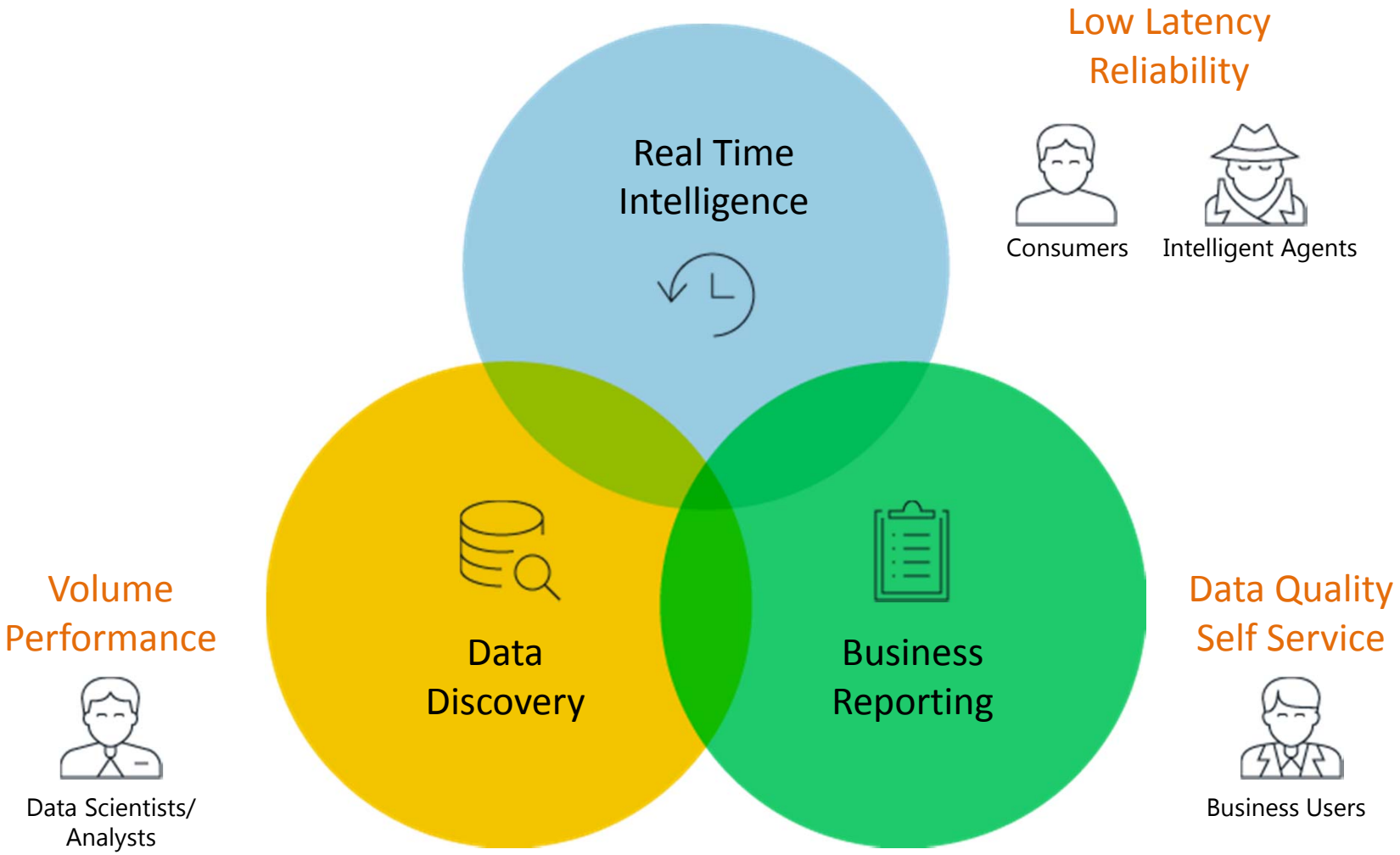
Machine Log Data
Application logs, event logs, server data, CDRs, clickstream data etc.

Sensor Data
Smart electric meters, medical devices, car sensors, road cameras etc.

Big Data Analytics

	Traditional Analytics (BI)	vs	Big Data Analytics
Focus on	• Descriptive analytics • Diagnosis analytics		• **Predictive analytics** • **Data Science**
Data Sets	• Limited data sets • Cleansed data • Simple models		• Large scale data sets • More types of data • Raw data • Complex data models
Supports	**Causation:** what happened, and why?		**Correlation**: new insight More accurate answers

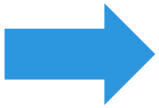
Big Data Analytics Use Cases



Big Data Analytics Reference Architectures

Architecture Drivers:

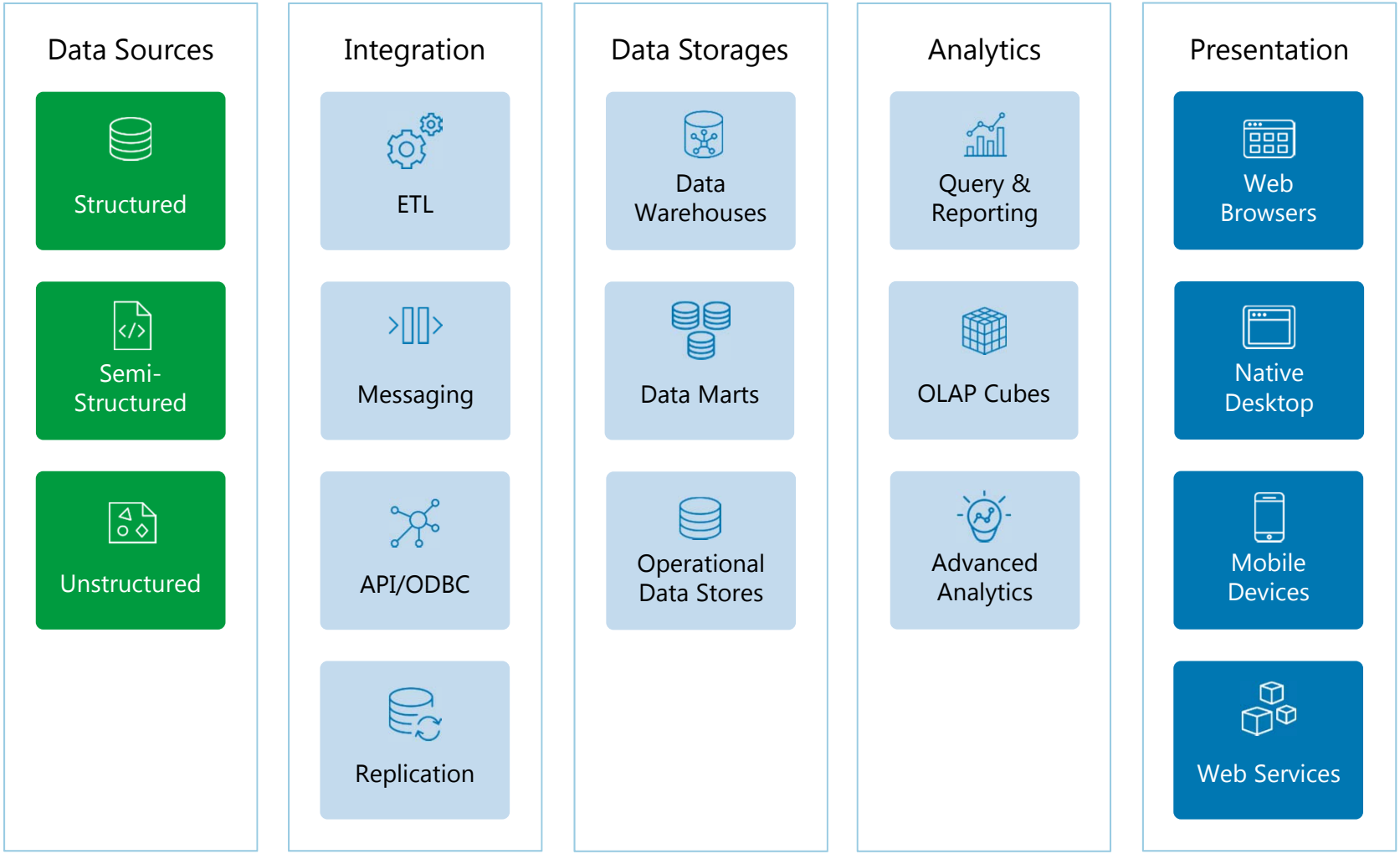
- Volume
- Sources
- Throughput
- Latency
- Extensibility
- Data Quality
- Reliability
- Security
- Self-Service
- Cost



Reference Architectures:

- Extended Relational
- Non-Relational
- Hybrid

Relational Reference Architecture

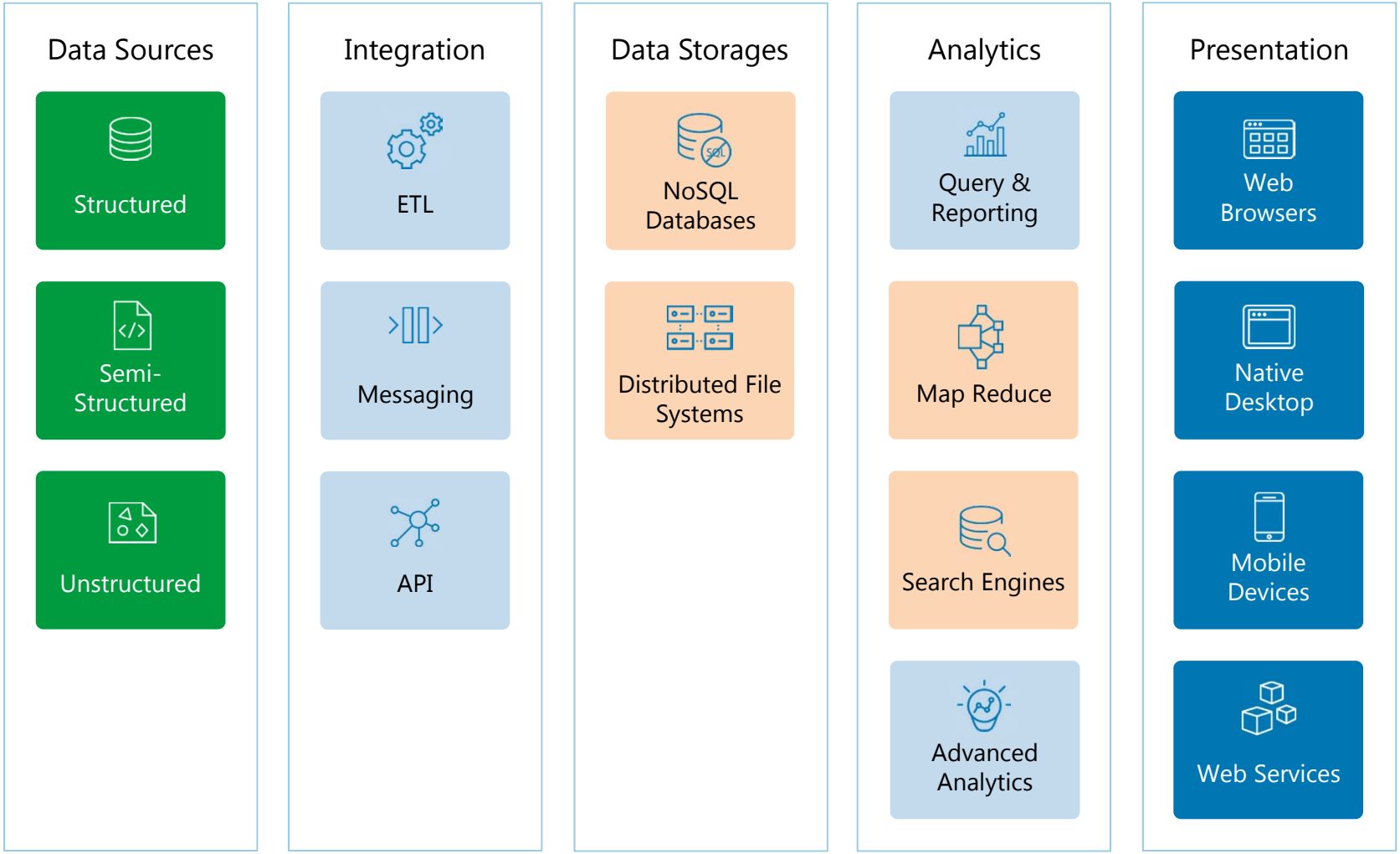


Extended Relational Reference Architecture

Data Sources	Integration	Data Storages	Analytics	Presentation
Structured	ETL	Data Warehouses	Query & Reporting	Web Browsers
Semi-Structured	Messaging	Data Marts	OLAP Cubes	Native Desktop
Unstructured	API/ODBC	Operational Data Stores	Advanced Analytics	Mobile Devices
	Replication			Web Services

Key components affected with Big Data challenges

Non-Relational Reference Architecture



Key components introduced with non-relational movement

Extended Relational vs. Non-Relational Architecture

Architecture Drivers	Extended Relational	Non-Relational
Large data volume		
Self-service (ad-hoc reporting)		
Unstructured data processing		
High data model extensibility		
High data quality and consistency		
Extensive security		
Reliability and fault-tolerance		
Low latency (near-real time)		
Low cost		
Skills availability		

Extended Relational vs. Non-Relational Architecture

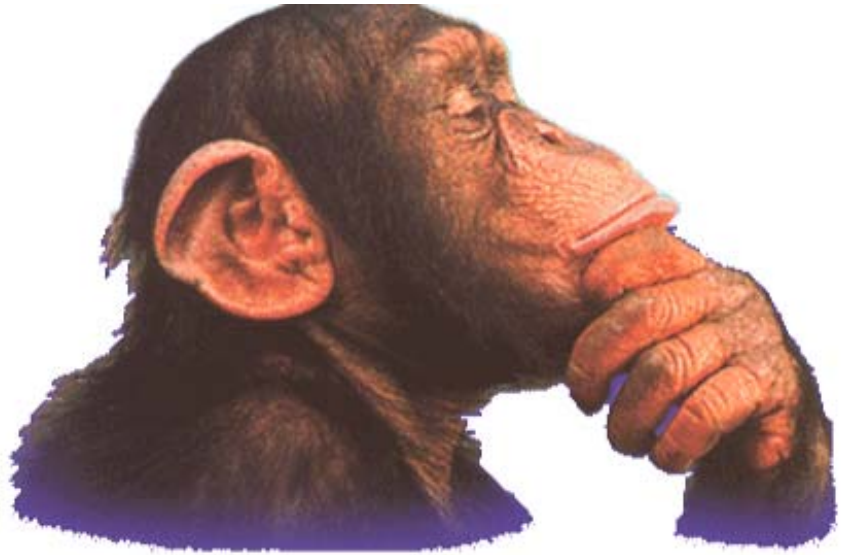
Architecture Drivers	Extended Relational	Non-Relational
Large data volume		
Self-service (ad-hoc reporting)		
Unstructured data processing		
High data model extensibility		
High data quality and consistency		
Extensive security		
Reliability and fault-tolerance		
Low latency (near-real time)		
Low cost		
Skills availability		

Extended Relational vs. Non-Relational Architecture

Architecture Drivers	Extended Relational	Non-Relational
Large data volume		
Self-service (ad-hoc reporting)		
Unstructured data processing		
High data model extensibility		
High data quality and consistency		
Extensive security		
Reliability and fault-tolerance		
Low latency (near-real time)		
Low cost		
Skills availability		

Relational vs. Non-Relational Architecture

Relational



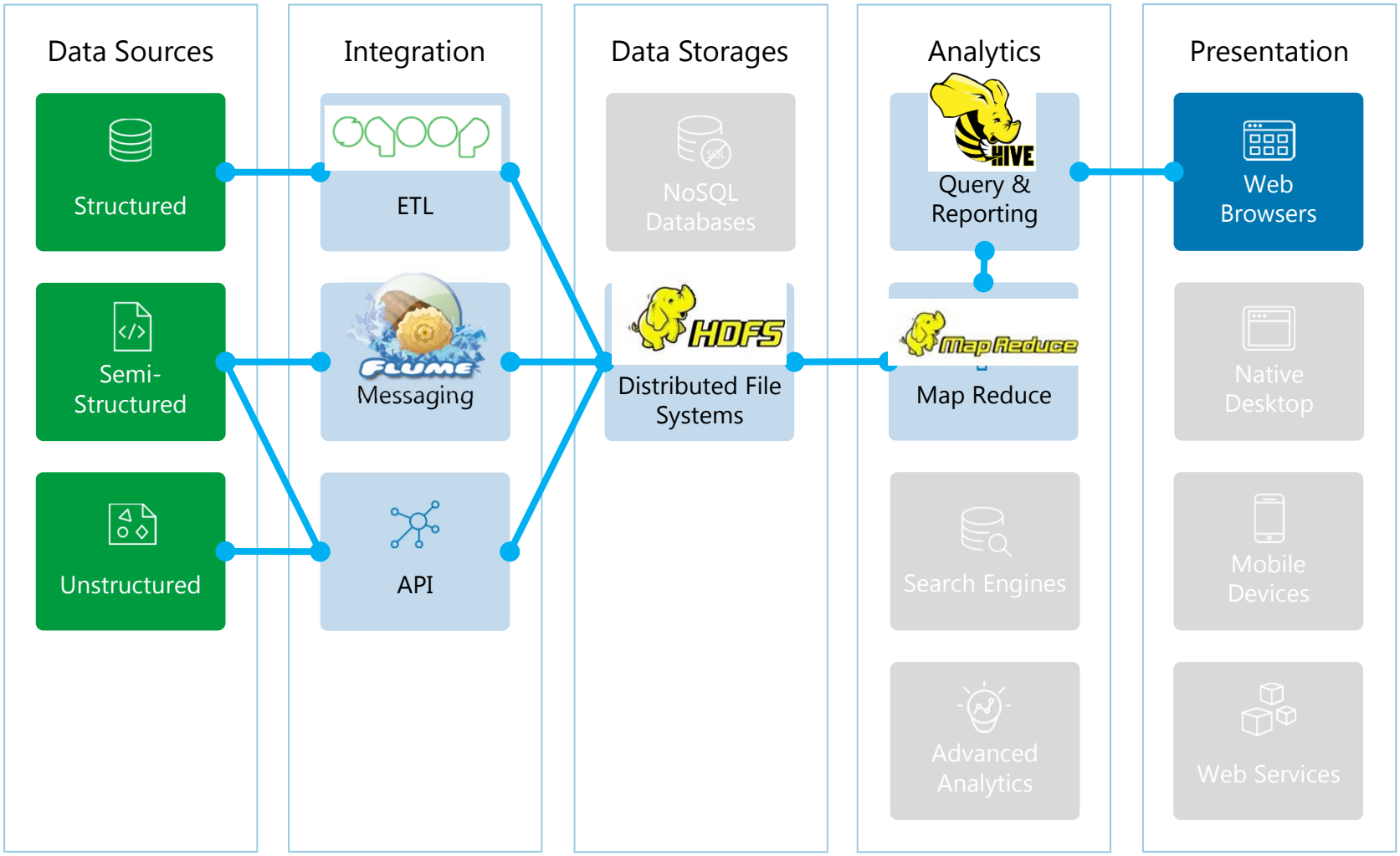
- Rational
- Predictable
- Traditional

Non-Relational

- Agile
- Flexible
- Modern

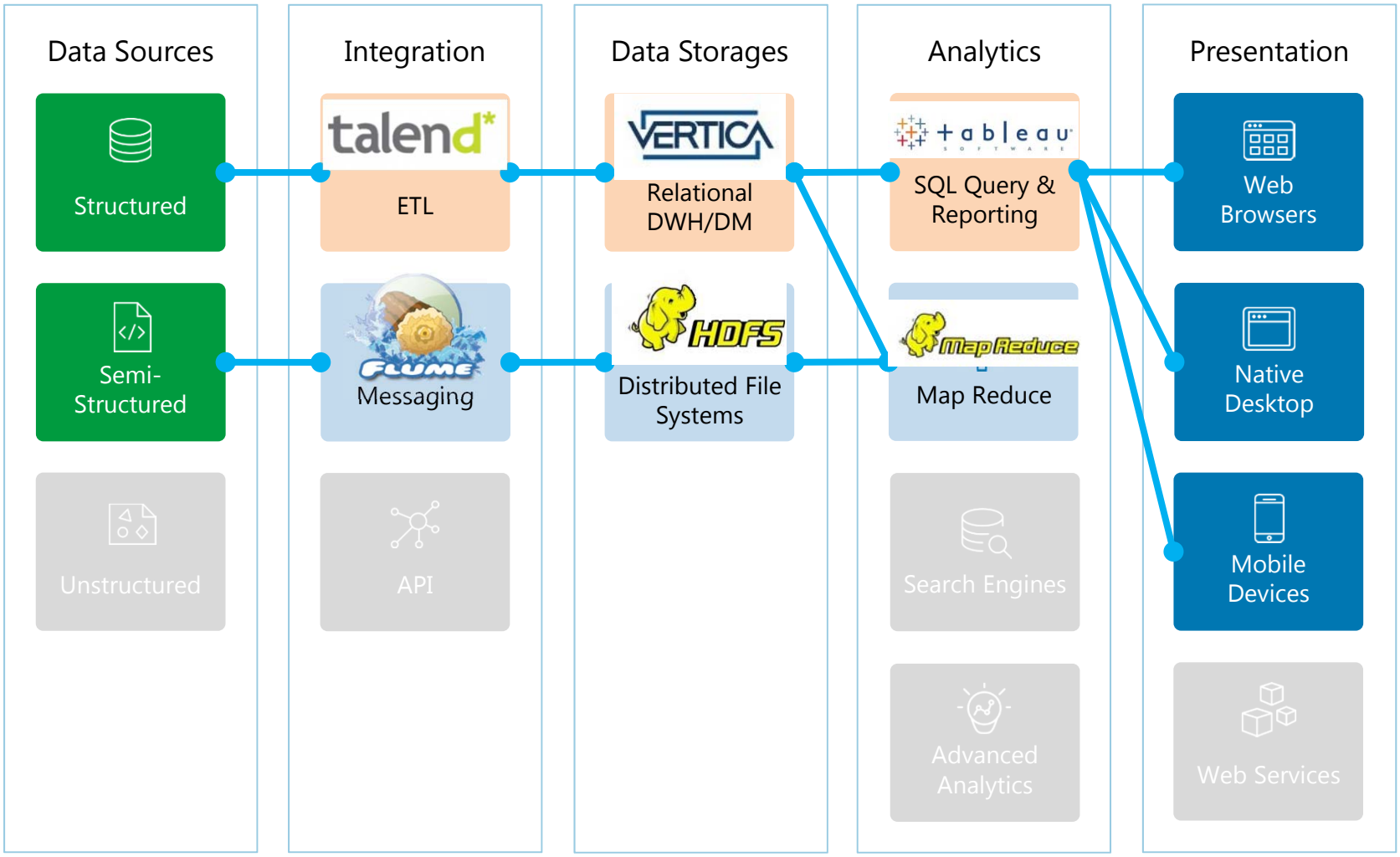
Big Data Analytics Use Cases

Data Discovery: Non-Relational Architecture



Big Data Analytics Use Cases

Business Reporting: Hybrid Architecture

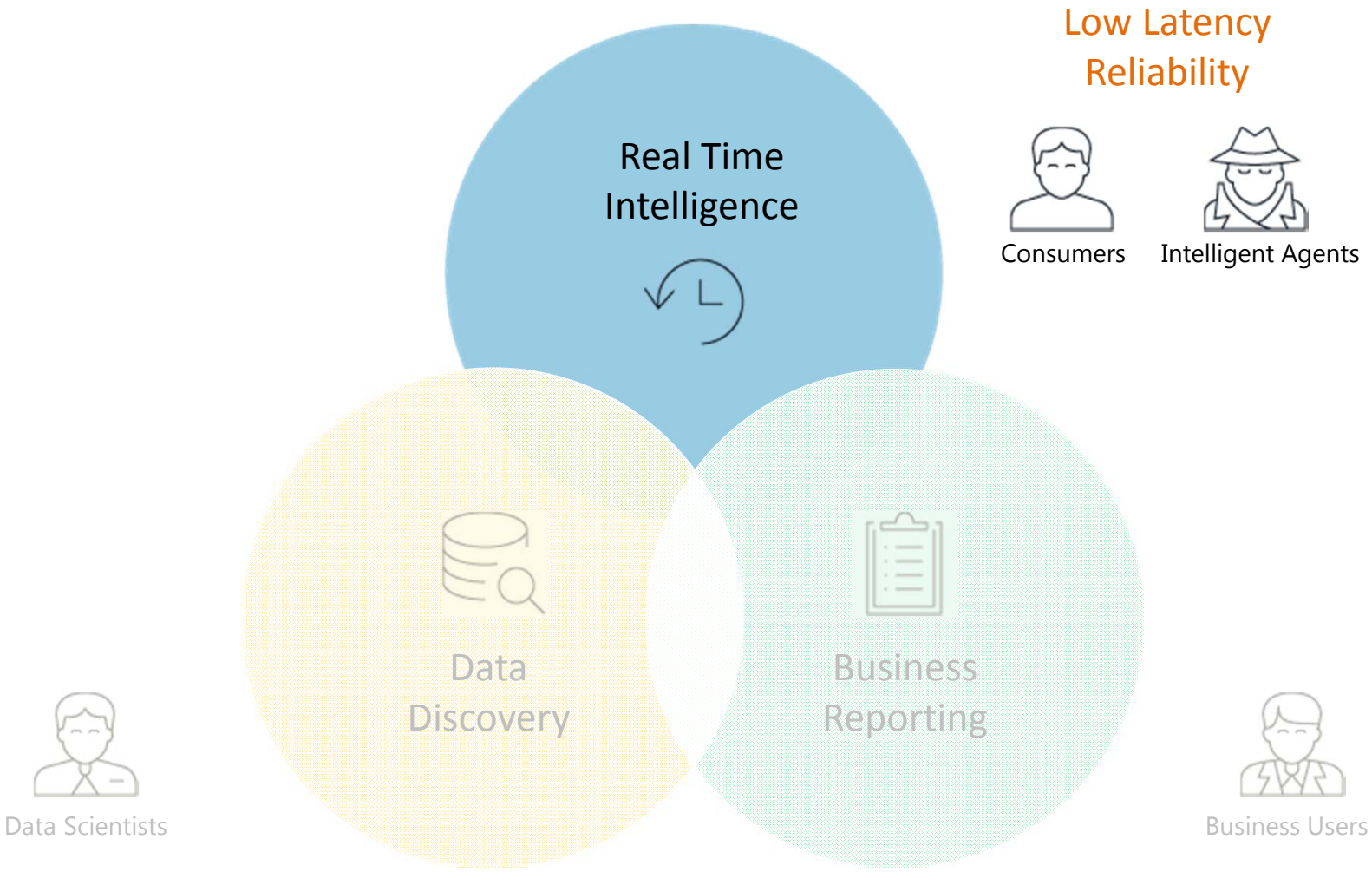


SoftServe

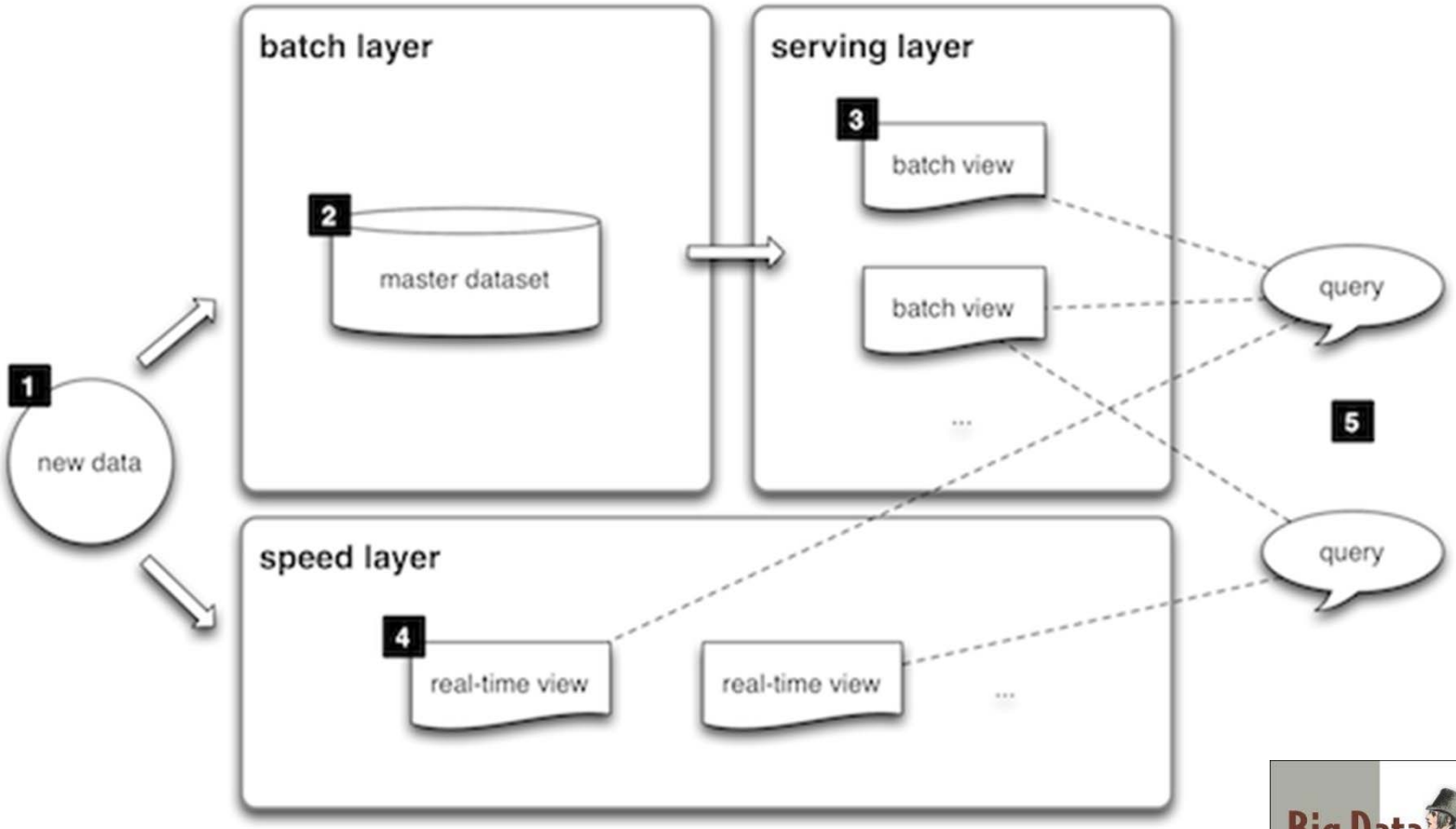
Extended Relational components

Non-relational components

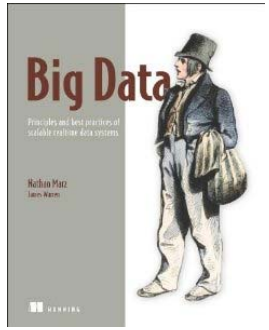
Big Data Analytics Use Cases

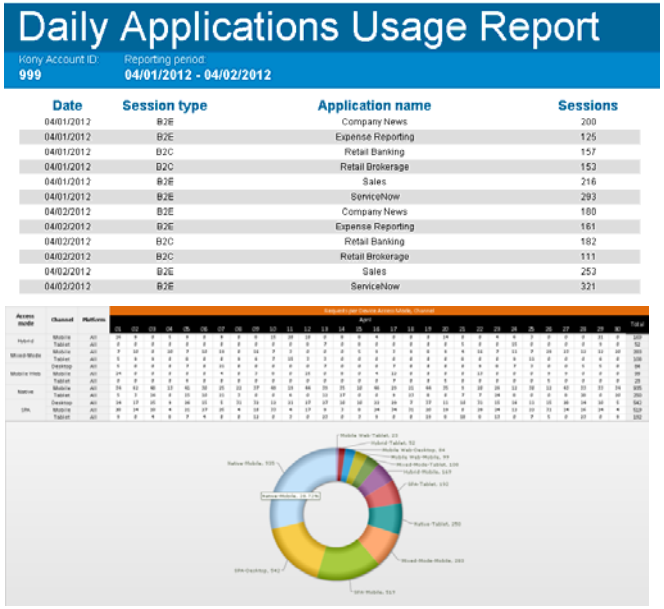


Lambda Architecture



Source:





Case Study #1: Usage & Billing Analysis

Business Goals:

✓ Provide visual environment for building custom mobile application

✓ Charge customers based on the platform they are using, number of consumers' applications etc.

Business Area:

Cloud based platform for building, deploying, hosting and managing of mobile applications

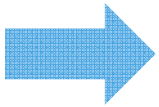
Architectural Decisions

Architecture Drivers:

- Volume (> 10 TB)
- Sources (Semi-structured - JSON)
- Throughput (> 10K/sec)
- Latency (2 min)
- Extensibility (Custom metrics)
- Data Quality (Consistency)
- Reliability (24/7)
- Security (Multitenancy)
- Self-Service (Ad-Hoc reports)
- Cost (The less the better ☺)
- Constraints (Public Cloud)

Trade-off:

	Extended Relational	Non-Relational
Extensibility	-	+
Data Quality	+	-
Self-Service	+	-

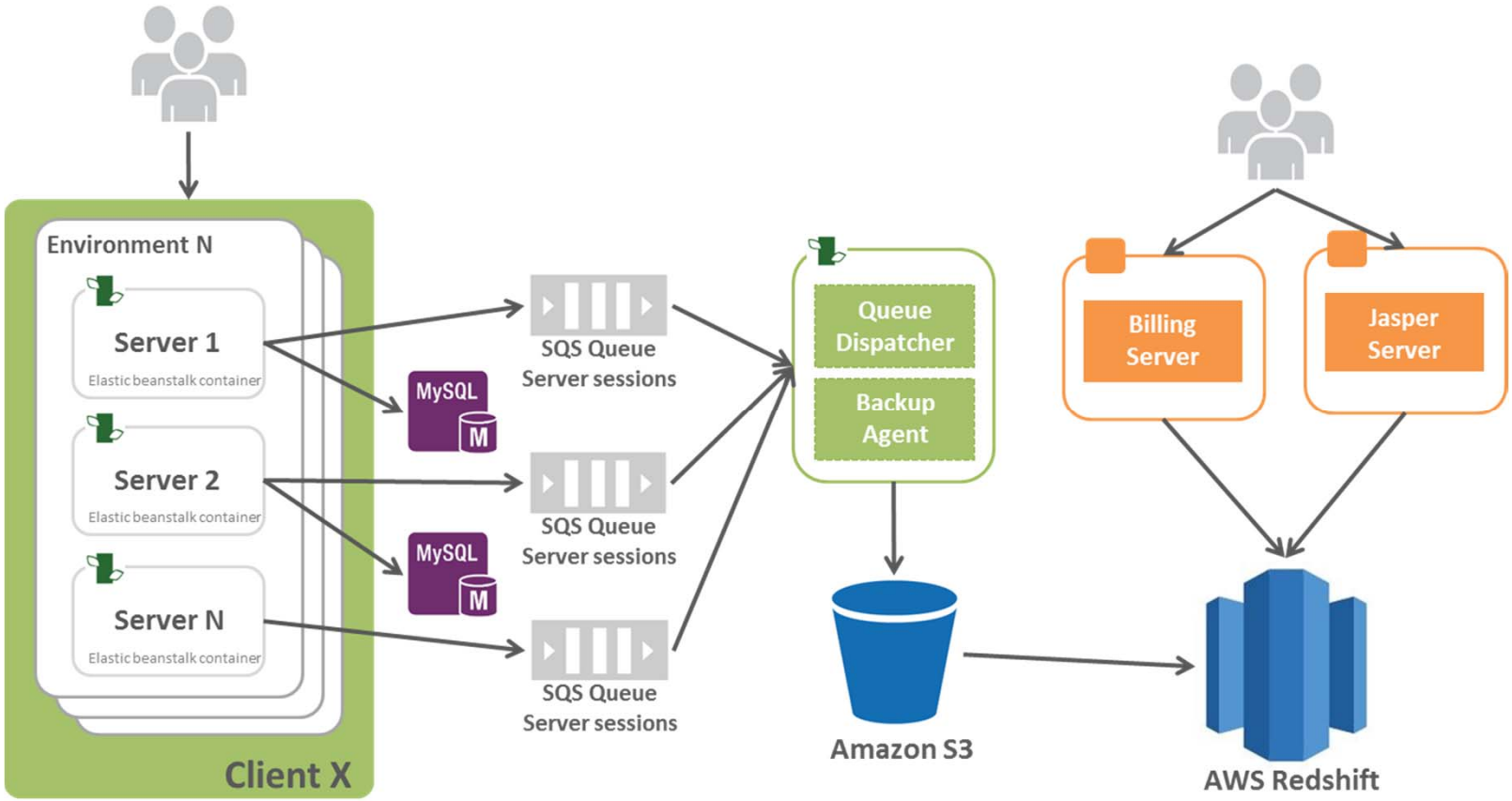


✓ **Extended Relational Architecture**
✓ Extensibility via **Pre-allocated Fields** pattern

Solution Architecture

Technologies:

- Amazon Redshift
- Amazon SQS
- Amazon S3
- Elastic Beanstalk
- Jaspersoft BI Professional
- Python



Case Study #2: Clickstream for retail website

Business Goals:

✓ Build in-house Analytics Platform for ROI measurement and performance analysis of every product and feature delivered by the e-commerce platform;
✓ Provide the ability to understand how end-users are interacting with service content, products, and features on sites;
✓ Do clickstream analysis;
✓ Perform A/B Testing

Business Area:

Retail. A platform for e-commerce and collecting feedbacks from customers

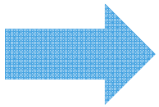
Architectural Decisions

Architecture Drivers:

- Volume (45 TB)
- Sources (Semi-structured - JSON)
- Throughput (> 20K/sec)
- Latency (1 hour)
- Extensibility (Custom tags)
- Data Quality (Not critical)
- Reliability (24/7)
- Security (Multitenancy)
- Self-Service (Canned reports, Data science)
- Cost (The less the better ☺)
- Constraints (Public Cloud)

Trade-off:

	Extended Relational	Non-Relational
Volume/Scalability	+/-	+
Throughput	+	+
Self-Service	+	+/-
Extensibility	-	+

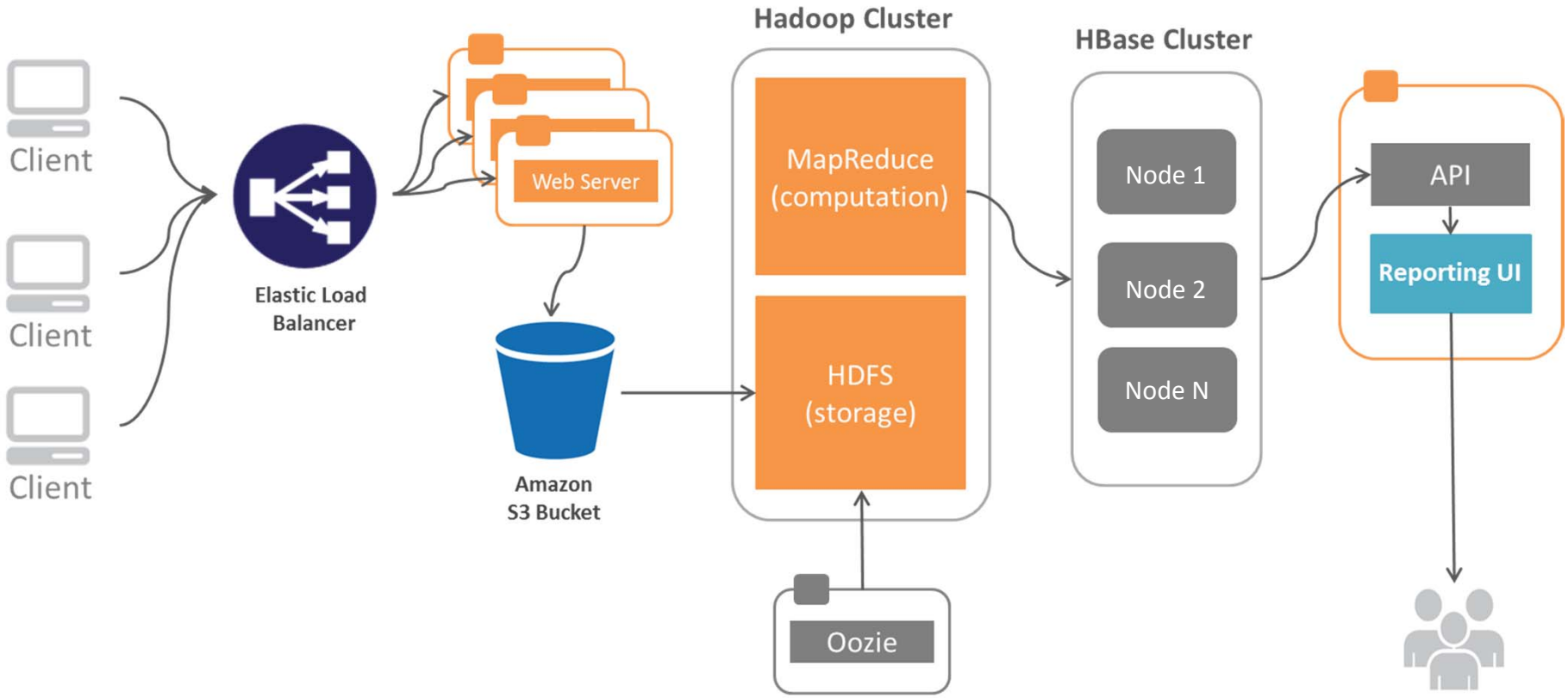


✓ **Non-Relational Architecture**
✓ Reporting via **Materialized View** pattern

Solution Architecture

Technologies:

- Amazon S3
- Flume
- Hadoop/HDFS, MapReduce
- HBase
- Oozie
- Hive



Tips for Designing Big Data Solutions

- Understand data users and sources
- Discover architecture drivers
- Select proper reference architecture
- Do trade-off analysis, address cons
- Map reference architecture to technology stack
- Prototype, re-evaluate architecture
- Estimate implementation efforts
- Set up devops practices from the very beginning
- Advance in solution development through “small wins”
- Be ready for changes, big data technologies are evolving rapidly

SoftServe

- Leading global Product and Application Development partner founded in 1993
- 3,300+ employees across North America, Ukraine and Western Europe
- Thousands of successful outsourcing projects!

Clients include:

Allscripts, Cisco, Logitech, HP, Parker, IBM, ThermoFisher Scientific, bmcsoftware

SaaS/Cloud Solutions . Mobility Solutions . UX/UI
BI/Analytics/Big Data . Software Architecture . Security