

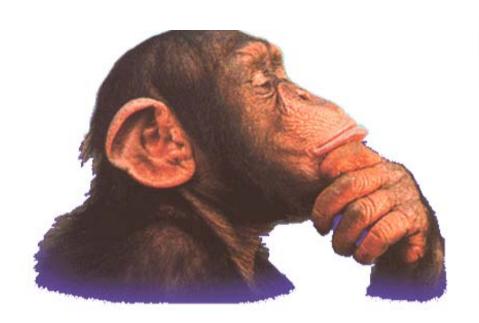


BIG DATA ANALYTICS

REFERENCE ARCHITECTURES AND CASE STUDIES

Relational vs. Non-Relational Architecture

Relational



- Rational
- Predictable
- Traditional

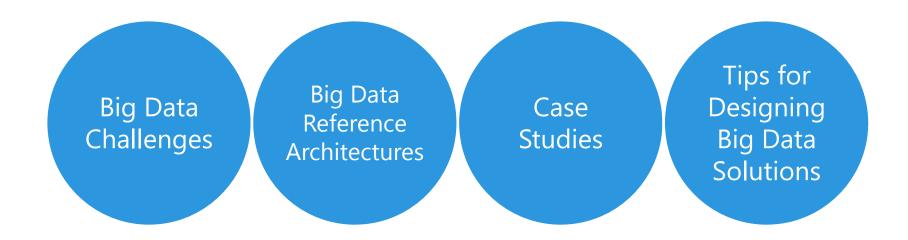
Non-Relational



- Agile
- Flexible
- Modern

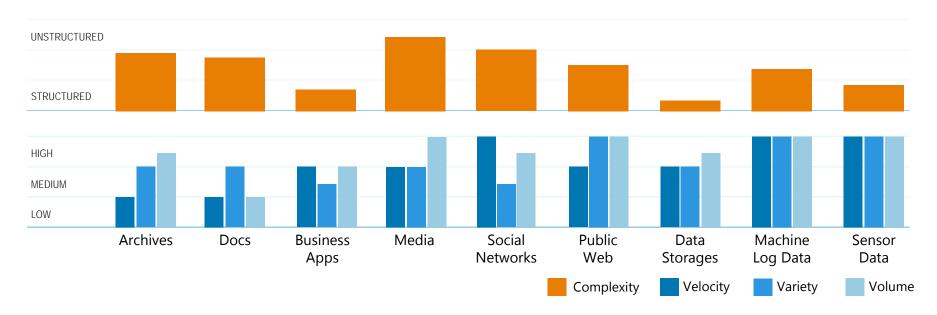


Agenda





Big Data Challenges





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.

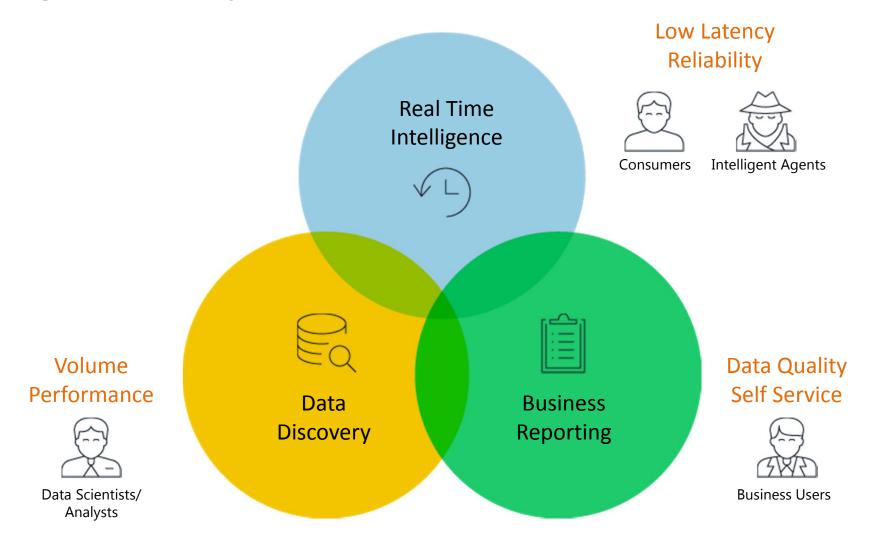


Big Data Analytics

	Traditional Analytics (BI)	vs Big Data Analytics
Focus on	Descriptive analyticsDiagnosis analytics	Predictive analyticsData Science
Data Sets	Limited data setsCleansed dataSimple 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

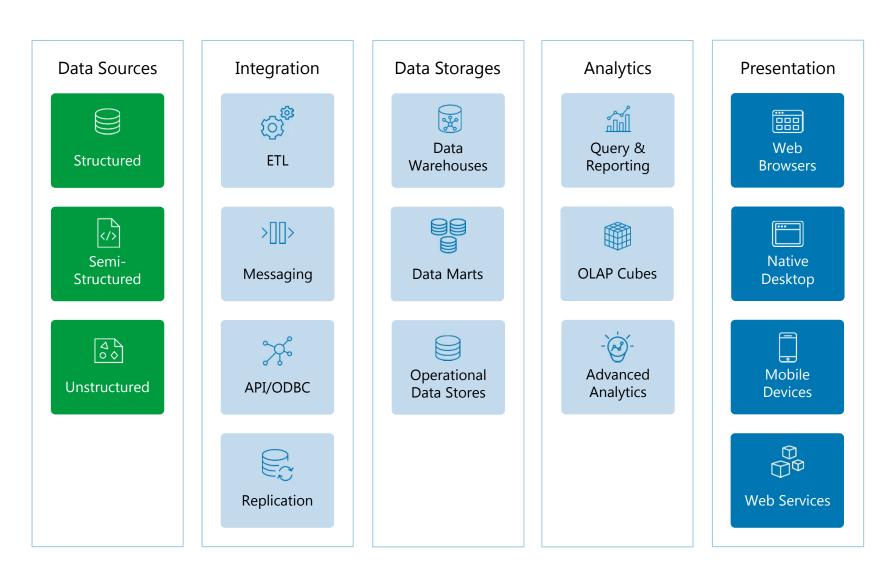


Reference Architectures:

- Extended Relational
- Non-Relational
- Hybrid



Relational Reference Architecture



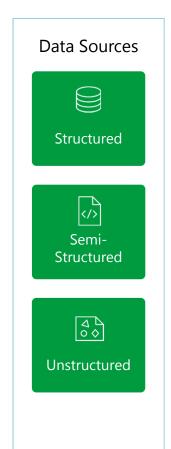


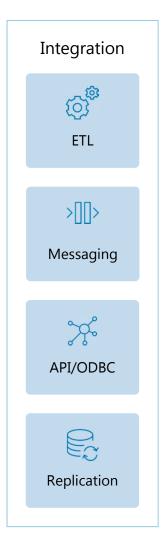
Extended Relational Reference Architecture

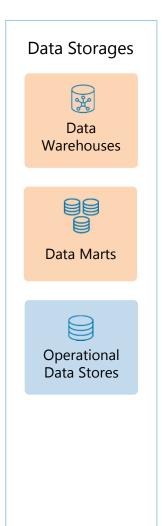












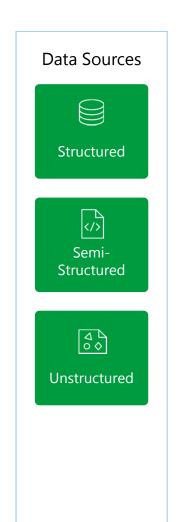


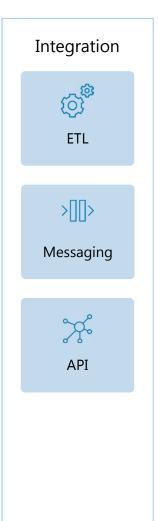


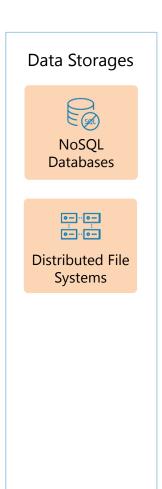




Non-Relational Reference Architecture















Extended Relational vs. Non-Relational Architecture

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



Extended Relational vs. Non-Relational Architecture

Architecture Drivers	Extended Relational	Non-Relational
Large data volume	OF	
Self-service (ad-hoc reporting)	Œ	
Unstructured data processing		I
High data model extensibility		I
High data quality and consistency	OF	
Extensive security	OF	
Reliability and fault-tolerance	OF	o É
Low latency (near-real time)	+	4
Low cost		E
Skills availability	OF	



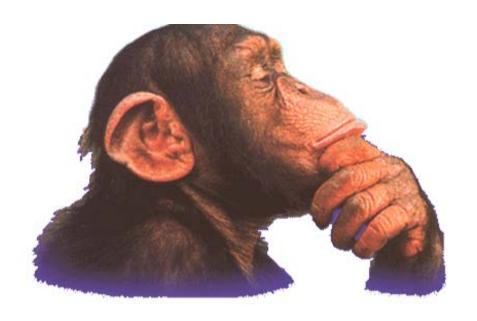
Extended Relational vs. Non-Relational Architecture

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



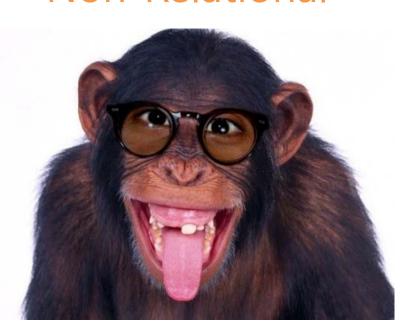
Relational vs. Non-Relational Architecture

Relational



- Rational
- Predictable
- Traditional

Non-Relational



- Agile
- Flexible
- Modern

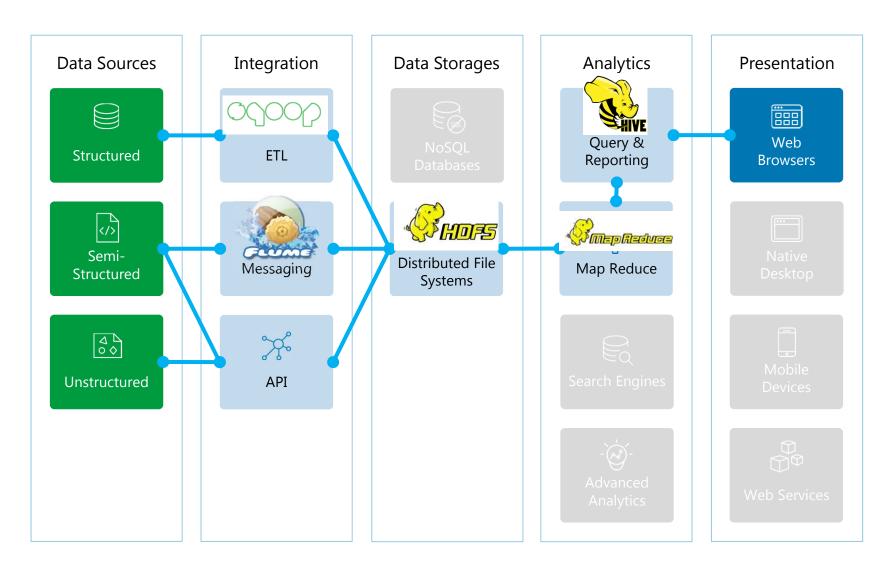


Big Data Analytics Use Cases





Data Discovery: Non-Relational Architecture



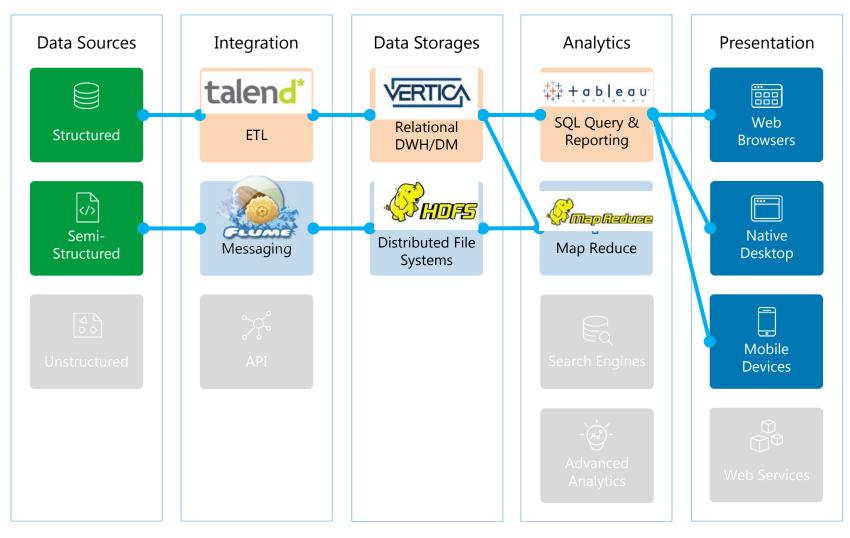


Big Data Analytics Use Cases





Business Reporting: Hybrid Architecture







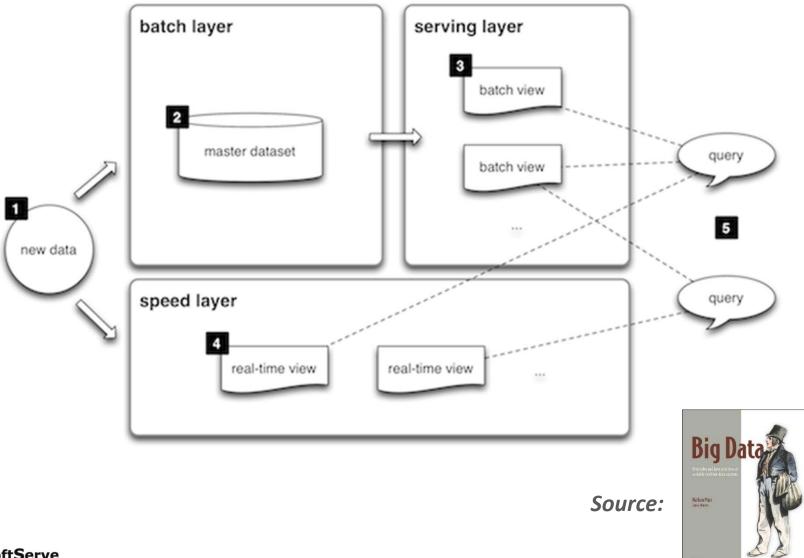


Big Data Analytics Use Cases





Lambda Architecture





999	Accour		Reporting period: 04/01/2012 - 04/02/2012																															
	Date	Э	Session type Application name																							าร								
-	04/01/20	112	B2E						B2E Company News														200											
	04/01/20					32E				Expense Reporting													125											
	04/01/20				_	12C												Ban					157											
	04/01/20					12C										Re		groko	rag	90								153						
	04/01/20					126												ales									216							
	04/01/20					32E												iceN										293						
	04/02/20					32E												any h										180						
	04/02/20		B2E Expense Reporting											161																				
	04/02/20		B2C Retail Banking										182																					
	04/02/20		B2C Retail Brokerage										111																					
	04/02/20 04/02/20		B2E Sales								253																							
	14102020	112	B2E SeniceNow 321																															
Access mode	Channel	Platform													Ees	ued i j	er Con Apr		0	sde, O	arnel												Fati	
Habital	More	All	26 26	1	03	1		8		1		15	21	10	7	1		1	,	ŗ		14	-	-	4	4	2	1	-	1	21		149	
Total Works	Tablet Miblie	AT AT	÷	10	:	-	÷	÷	÷	÷	1	÷	÷	1	;	÷	÷	1	i	÷	÷	÷		÷	÷	11	11	20	÷	-	10	10	52	
Bally You	Tablet Desktop Monte	AT	1	÷			3		11	÷	:	÷	7		1				:				÷	:	÷	•	,		÷	1			3/56 04 22	
AUGUS THES	Tablet	AT	p 24			12	÷	1 20	15	10	17	i	1	:		11	10	i	,	*	i	1 15		10	1	11	7	1	i	1	i		23 935	
Marrie	Tablet Desistop	AT	5	12	×	1	15	10	10	3	0 30	12	1		10	17			10	10	*	*	i	7 70	14	0	*	# 15		30	*	10	250 540	
584.	Muore Toblet	AT AT	,,	14	10	:	÷	17	×	;	10	10	;	12	9 20	1	ï	ř)= 0	11	30 20	12		10	12	12	10	111	14	16 22	14	;	525 150	
																Male		Table	Table	0.12	a de frança													
										Salive I	the below.	135 \		1		7			F101	te wat	-14/0.0	. 99												
			Nation (Holden, 1978) Water State of Holden, 1978 Water State of Holden, 1978 Water State of Holden, 1978 Water State of Holden, 1978																															
														. 10.77	5		V		M	/ SEA	Tablet.	192												
														1					١			ene. z												
												١		١			1		7	-41	etion-Ti	ana, s	50											



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 ②)
- Constraints (Public Cloud)

-			CC
	r 0 c	10	·off:
	1 40	10-	-()
	IUC		\sim 11.

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



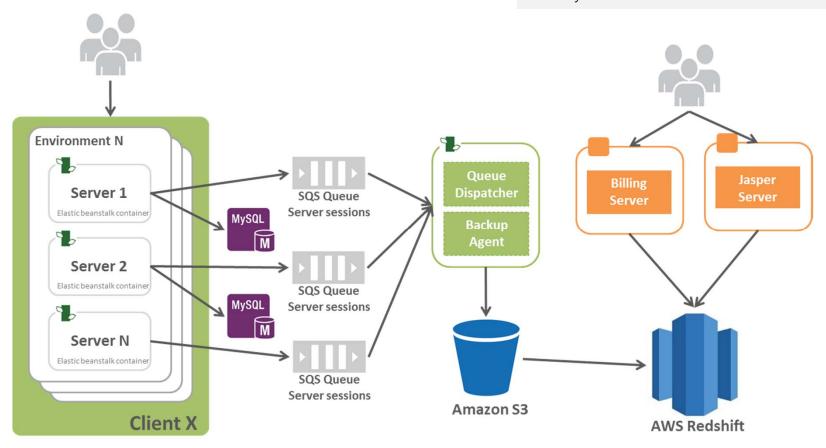
- ✓ Extended Relational Architecture
- ✓ Extensibility via Pre-allocated Fields pattern



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 ②)
- Constraints (Public Cloud)

Trade-off:

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



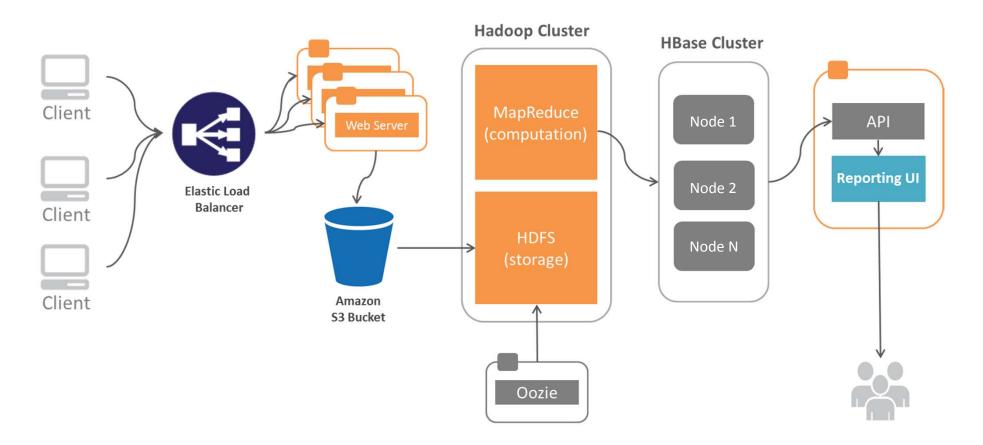
- ✓ Non-Relational Architecture
- ✓ Reporting via Materialized View pattern



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





- 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!



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



Thank You!

SoftServe US Office

One Congress Plaza, 111 Congress Avenue, Suite 2700 Austin, TX 78701

Tel: 512.516.8880

Contacts

Serhiy Haziyev: shaziyev@softserveinc.com

Olha Hrytsay: ohrytsay@softserveinc.com

