IS04 – Analytics on Linux on IBM z Systems

Arwed Tschoeke, Client Center Böblingen

arwed.tschoeke@de.ibm.com

25. October 2016

Analytics with Linux on z systems

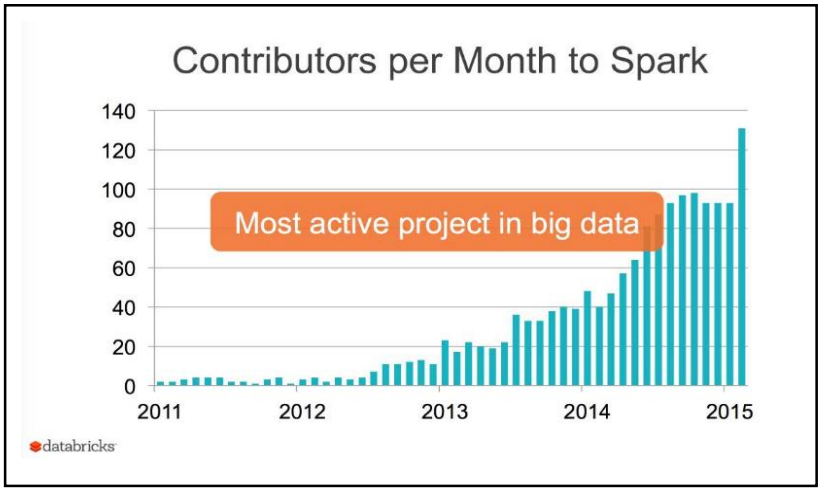
- New Kids on the Block – Spark and Hadoop
- Cognitive Analytics – Watson Explorer

What is Spark?

- An Apache Foundation open source project; not a product
- An in-memory compute engine that works with data; not a data store
- Enables highly iterative analysis on large volumes of data at scale
- Unified environment for data scientists, developers and data engineers
- Radically simplifies the process of developing intelligent apps fueled by data

Brief History of Spark

2014 – Apache Spark top-level
2014 – 1.2.0 release in December
2015 – 1.3.0 release in March
2015 – 1.4.0 release in June
2015 – 1.5.0 release in September
2016 – 1.6.0 release in January



IBM's Commitment to Spark:

Founding Member of AMPLab

Establish Spark Technology Center

Contributing to the core

Open Source SystemML

Educate one million data professionals

Key reasons for interest in Spark

Beware of the hype!

Performant

- In-memory architecture greatly reduces disk I/O
- Anywhere from **20-100x faster** for common tasks

Productive

- **Concise and expressive syntax**, especially compared to prior approaches
- **Single programming model** across a range of use cases and steps in data lifecycle
- **Integrated with common programming languages** – Java, Python, Scala
- **New tools** continually reduce skill barrier for access (e.g. SQL for analysts)

Leverages existing investments

- Works well within **existing Hadoop ecosystem**

Improves with age

- **Large and growing community** of contributors continuously improve full analytics stack and extend capabilities

Hadoop Advantages

Unlimited Scale

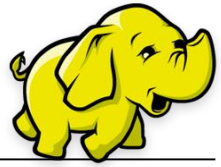
- Multiple data sources
- Multiple applications
- Multiple users

- Reliability
- Resiliency
- Security

Enterprise Platform

Wide Range of Data Formats

- Files
- Semi-structured
- Databases



Hadoop MapReduce Challenges

- Need deep Java skills
- Few abstractions available for analysts

Ease of Development

In-Memory Performance

- No in-memory framework
- Application tasks write to disk with each cycle

- Only suitable for batch workloads
- Rigid processing model

Combine Workflows

Spark Advantages

- Easier APIs
- Python, Scala, Java

Ease of Development

In-Memory Performance

- Resilient Distributed Datasets
- Unify processing

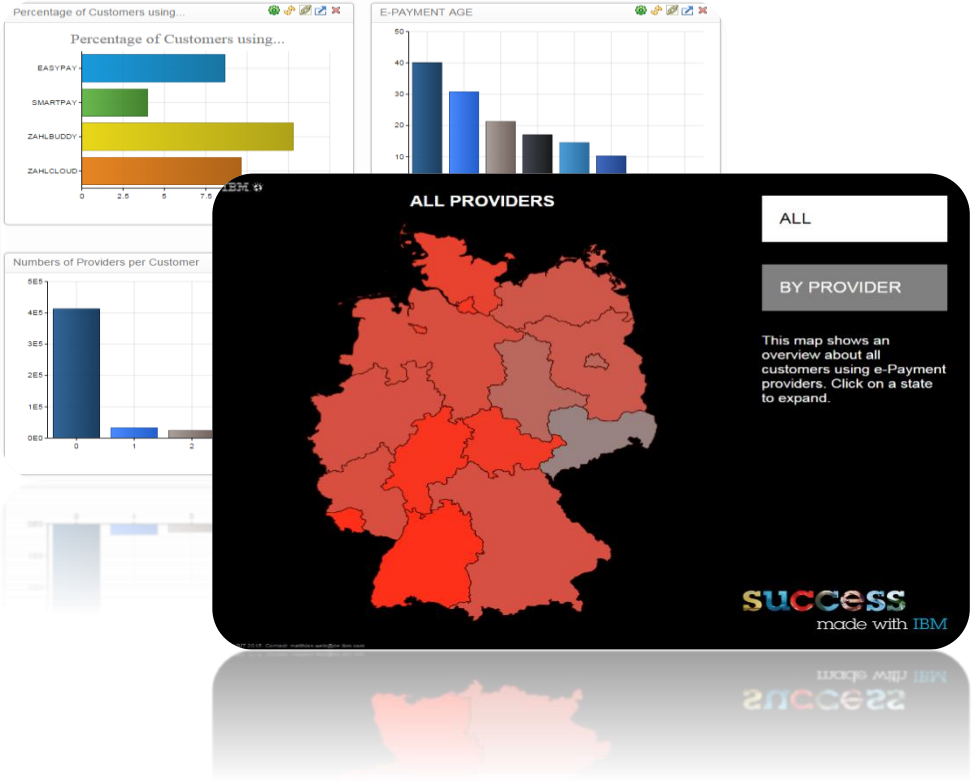
- Batch
- Interactive
- Iterative algorithms
- Micro-batch

Combine Workflows

Business Analytics: Usage of e-payment providers with Hadoop

Data generation facts

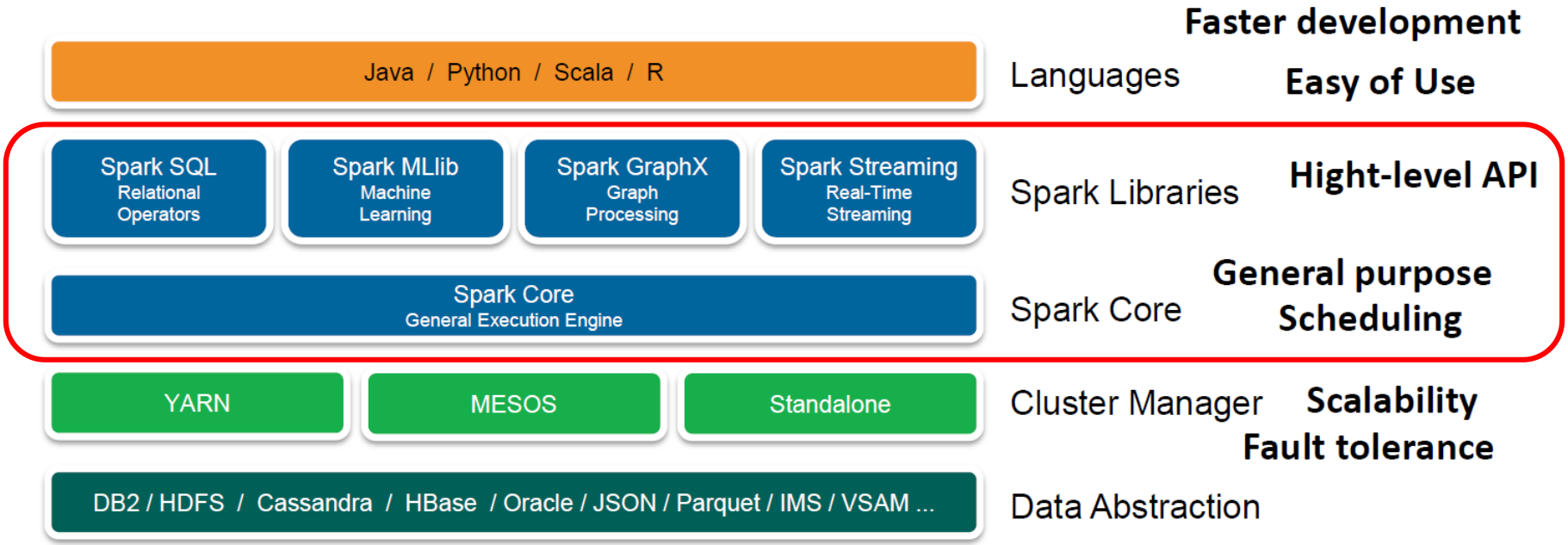
- 500,000 customers with address and birthday
- > 40 million transactions with amount, date and reference



Information of interest:

- Identify security vulnerabilities
- Market analysis
  - To establish a new e-payment method?
- Risk analysis
  - Can I lose customers by extending the e-payment market?
  - Lost sales in payments?
- Customer analysis

The Spark Stack, Architectural Overview

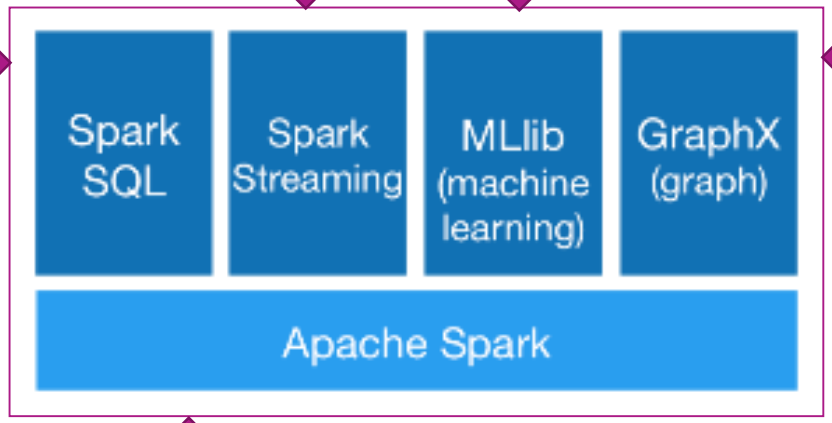


Apache Spark Runtime

Spark Streaming:
- Enables scalable, high-throughput processing of live data streams
- Live stream 'chopped' into batches based on time window

Spark MLIB
- Provides scalable machine learning library, has common machine learning functions
- Provides classification, regression, clustering, filtering, etc.

Spark SQL:
- Provides capability to perform relational queries via SQL (subset of HiveQL)
- Mix SQL queries with Spark applications



from http://spark.apache.org

Spark GraphX
- Spark APIs for graph style processing and iterative graph computations

Spark Core:
- Foundation providing task dispatching, scheduling, i/o
- Representation of Spark's basic unit of data: RDD

Cognitive Analytics – Watson Explorer

What is Cognitive Computing?

- Cognitive systems are able to learn their behavior through *education*
- That supports forms of *expression* that are more natural for human interaction
- Whose primary value is their *expertise*; and
- That continue to *evolve* as they experience new information, new scenarios, and new responses

… and does so at enormous scale.

1. Observe 2. Interpret 3. Evaluate 4. Decide

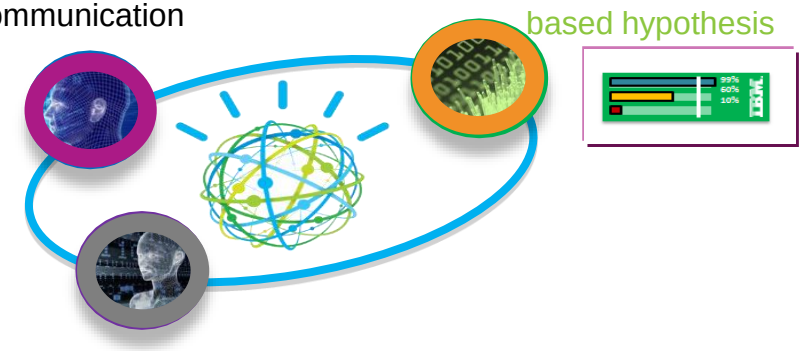
What is Watson?

Watson combines transformational capabilities

Delivering a new world experience, a Cognitive experience

1 Understands natural language and human style communication

2 Generates and evaluates evidence-based hypothesis



3 Adapts and learns from training, interaction, and outcomes

Watson:

- Understands me
- Engages me
- Learns and improves over time
- Helps me discover
- Establishes trust
- Has endless capacity for insight
- Operates in a timely fashion

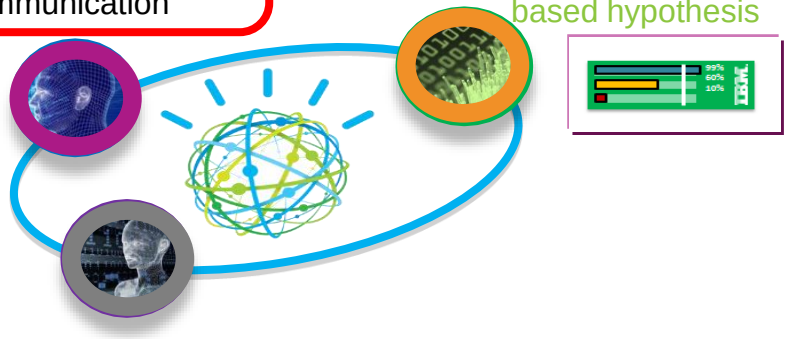
What is Watson Explorer?

Watson combines transformational capabilities

Delivering a new world experience, a Cognitive experience

1 Understands natural language and human style communication

2 Generates and evaluates evidence-based hypothesis



3 Adapts and learns from training, interaction, and outcomes

Watson:

- Understands me
- Engages me
- Learns and improves over time
- Helps me discover
- Establishes trust
- Has endless capacity for insight
- Operates in a timely fashion

Challenges

Watson Explorer

Information Access
Data, applications and services distributed on-premise and in cloud—employees struggle to get a complete view

Explore
Unified view of information from ALL sources to enable new insights and better decisions

Unstructured Content
80% of data is unstructured but only a small percentage leveraged for insights

Analyze
Delivers insights from unstructured content

Scaling Expertise
Pressure to increase performance and innovation—while doing more with less

Interpret
Applies cognitive computing to scale human expertise

Watson Explorer makes data from enterprise and non-enterprise silos easily accessible at “the point of impact” to people when they need it

Unstructured Content

Collaboration

Web

Email

File Systems

Content Management

Cloud

More…

Structured Data

Databases

Data Warehouses

Web Services

Cloud

More…

Organizations face major challenges when it comes to understanding their unstructured information…

- Volume, variety, velocity and veracity of information
- Inability to analyze and use unstructured data
- Difficulty analyzing and revealing patterns in data
- Manual, inefficient data analysis
- Siloed, fragmented and unknown information
- Inability to find and share data
- Inability to understand customer sentiment and preferences

Watson Explorer's content analytics creates actionable data from unstructured content

Unstructured content

Commander **4.0 Cu. Ft.** **26-Cycle King-Size** **washer** – **white**. I hate this machine. Have had 3 calls on machine. You can't wash **large items**, Won't clean in the middle. **Leaves dry spots** through the clothes, I can only do **½ basket** of clothes. Will **not clean** or **mix bleach** in with the water...

Watson Explorer
Deep natural-language analysis

Structured data for analysis

Product	**Commander**
Category	**4.0 Cu. Ft.**
Size	**26-Cycle King-Size**
Model	**washer**
Color	**white**
Issue	**large items**
Issue	**leaves dry spots**
Issue	**½ basket**
Issue	**not clean**
Issue	**mix bleach**

Watson Content Analytics provides the “why” behind the “what”

What is happening? → Why is it happening?

Analyzing structured data only gives you a partial view of the world around you

Only 20 percent of enterprise content is structured

Data analytics gives you the who, what, where and when of a subject

Mining unstructured content gives you a comprehensive understanding of the world around you

80 percent of enterprise content is unstructured

Content analytics distinctively adds the why and the how

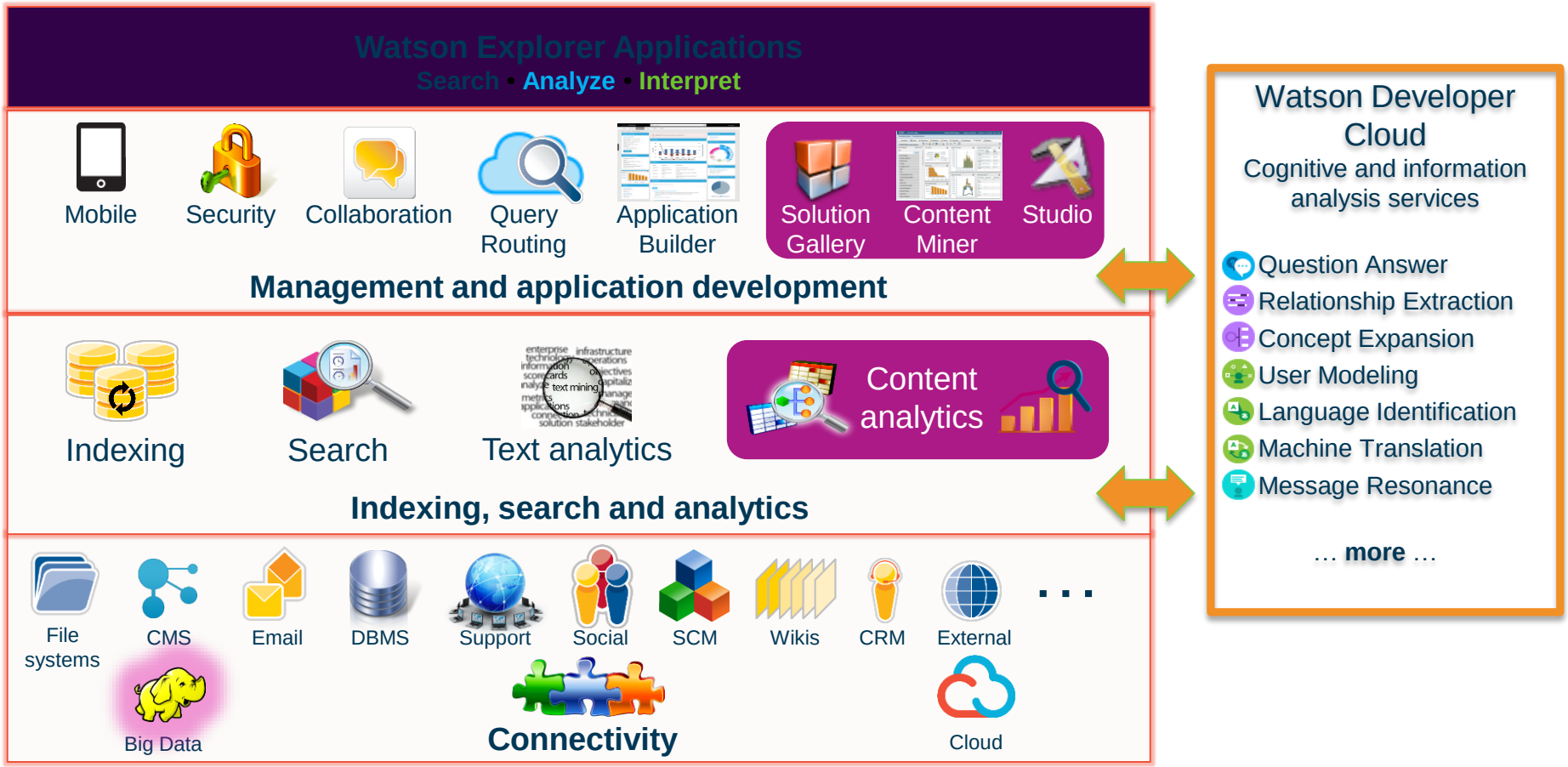
Use Watson Developer Cloud services to add cognitive capabilities to Watson Explorer applications

A growing list of services available*

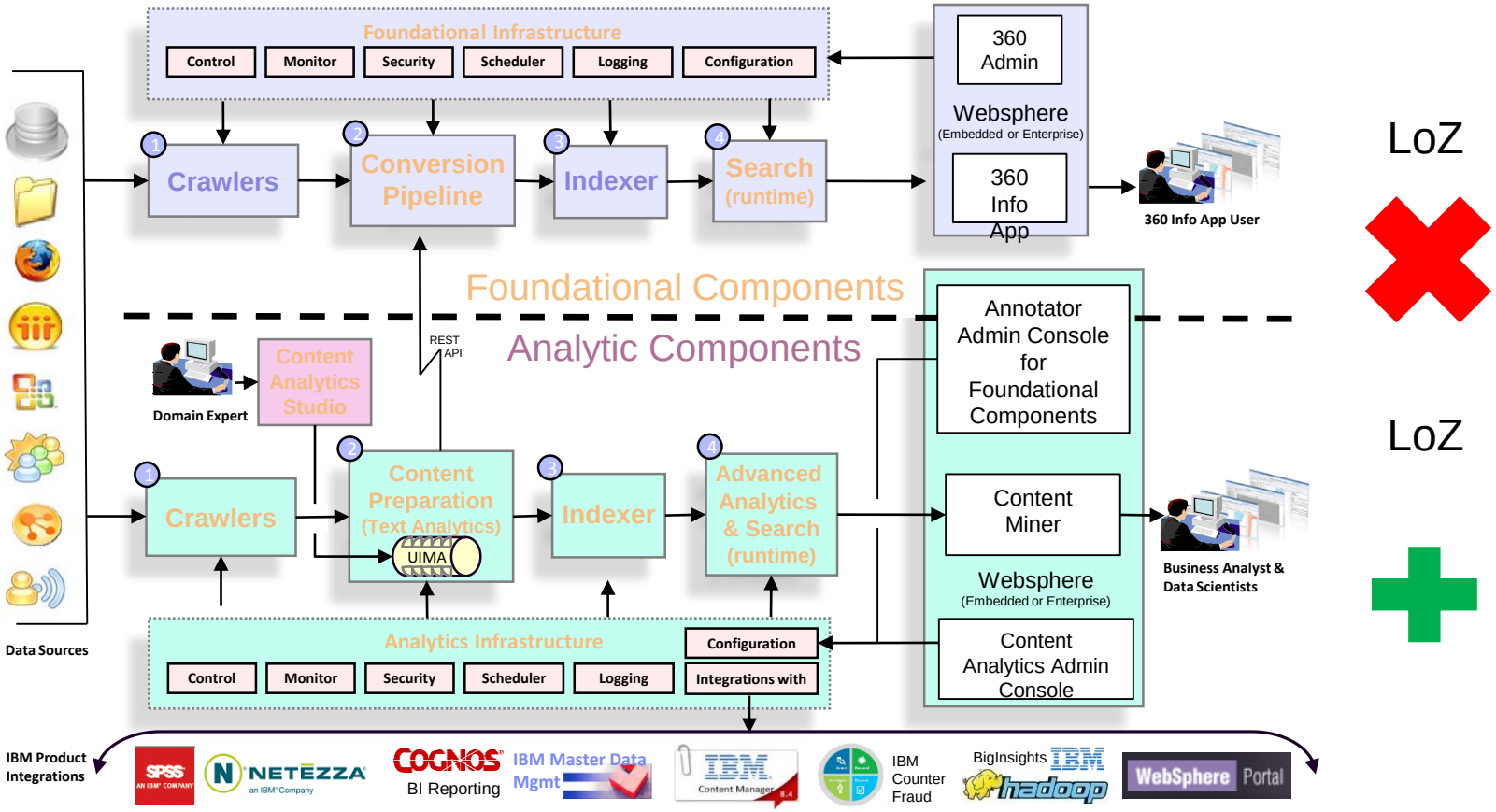
- AlchemyLanguage
- AlchemyVision**
- Language Translation
- Personality Insights
- Relationship Extraction
- Question and Answer
- Message Resonance
- Concept Expansion
- Tradeoff Analytics

*Services licensed separately. **Integrated with Watson Explorer Advanced Edition content analytics components. Some services may be in beta.

Watson Explorer Component View



Watson Explorer Foundational and Analytics Components



Why Watson Explorer on z?

- Don‘t move your z Systems sensitive data for Analysis

  ➔ Bring Watson Explorer there where your sensitive data reside

- Combine structured data with documents, semi-structured and unstructured data on z Systems (e.g. on Linux) and with external data without having to move those data to z Systems.
- Time to Market: Very fast provisioning of Linux guest, installation and deployment
- High speed access to data residing on z Systems
- Granular resource management to optimize performance of annotators
- Scale up Architecture to adopt resources to requirements on demand

Thank You for Your Attention!
Questions?

Watson Content Analytics for your business

Customer Insight

- Customer experience
- Customer satisfaction and survey analysis
- Product and service quality
- Churn prediction
- Marketing campaign development and execution
- New revenue opportunities
- Product enhancements

Crime Analytics

- Community policing
- Investigation analytics
- Incident management
- Antigang initiatives
- Antiterrorism initiatives
- Cyber crime investigation

Healthcare

- Diagnostic assistance
- Clinical treatment
- Critical care intervention
- Research for improved disease management
- Fraud detection and prevention
- Voice of the patient
- Claims management
- Prevention of readmissions
- Patient discharge and follow-up care

Insurance

- Risk assessment
- Fraud detection
- Policy and underwriting analysis
- Claims analysis, payment validation and loss review
- Reserve trending and optimization
- 360-degree view of the customer

Finance

- Anti–money laundering
- Internet banking fraud
- Operational efficiency
- Risk management and compliance

Arwed Tschoeke

IBM Client Center –
Systems and Software –
z ATS
IBM Germany Lab

Schoenaicher Str. 220
D-71032 Boeblingen

Phone +49 (0) 171 863 7780

arwed.tschoeke@de.ibm.com