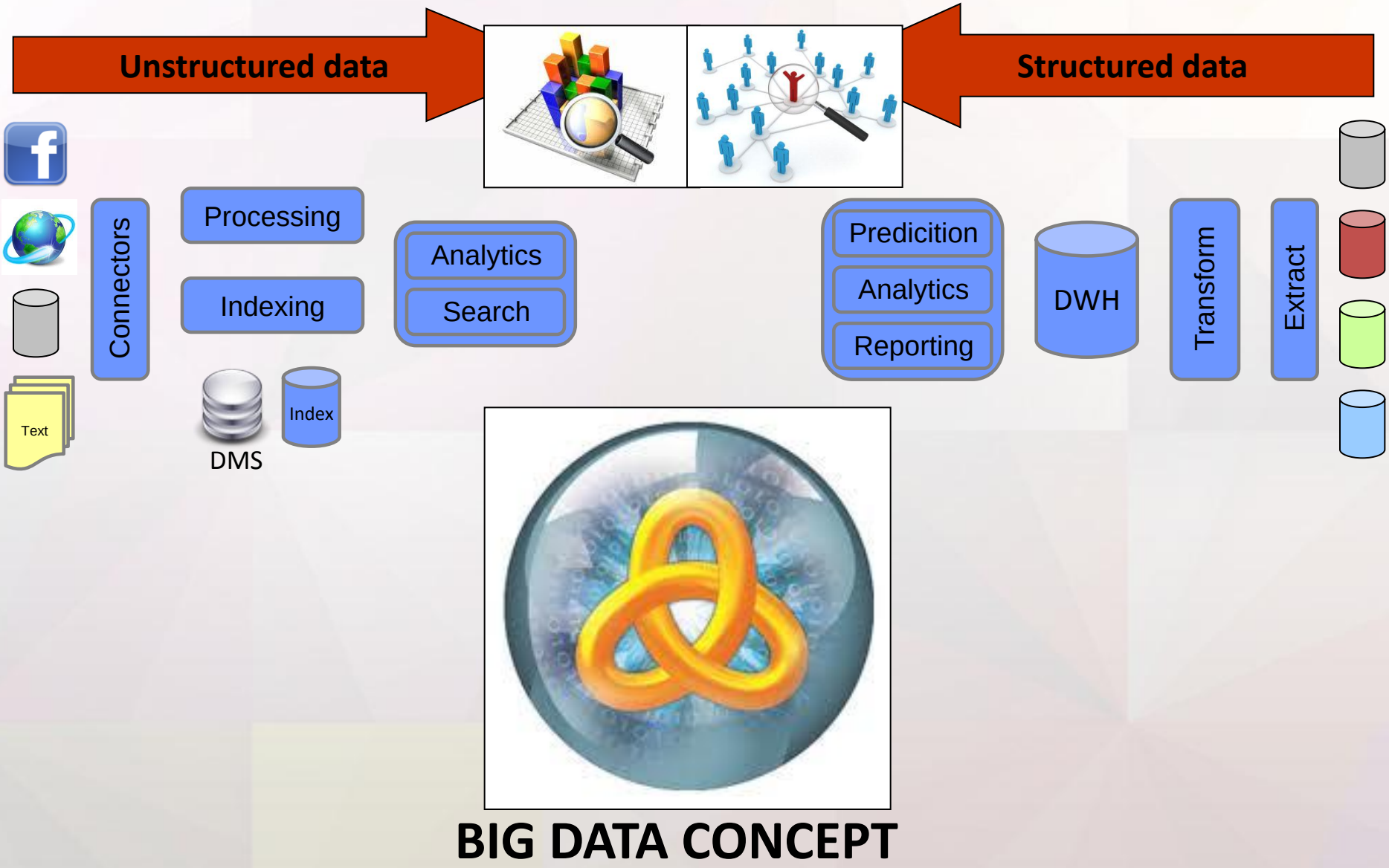
IBM BigData Analytics

SmarterAnalytics
CEE Competence Center

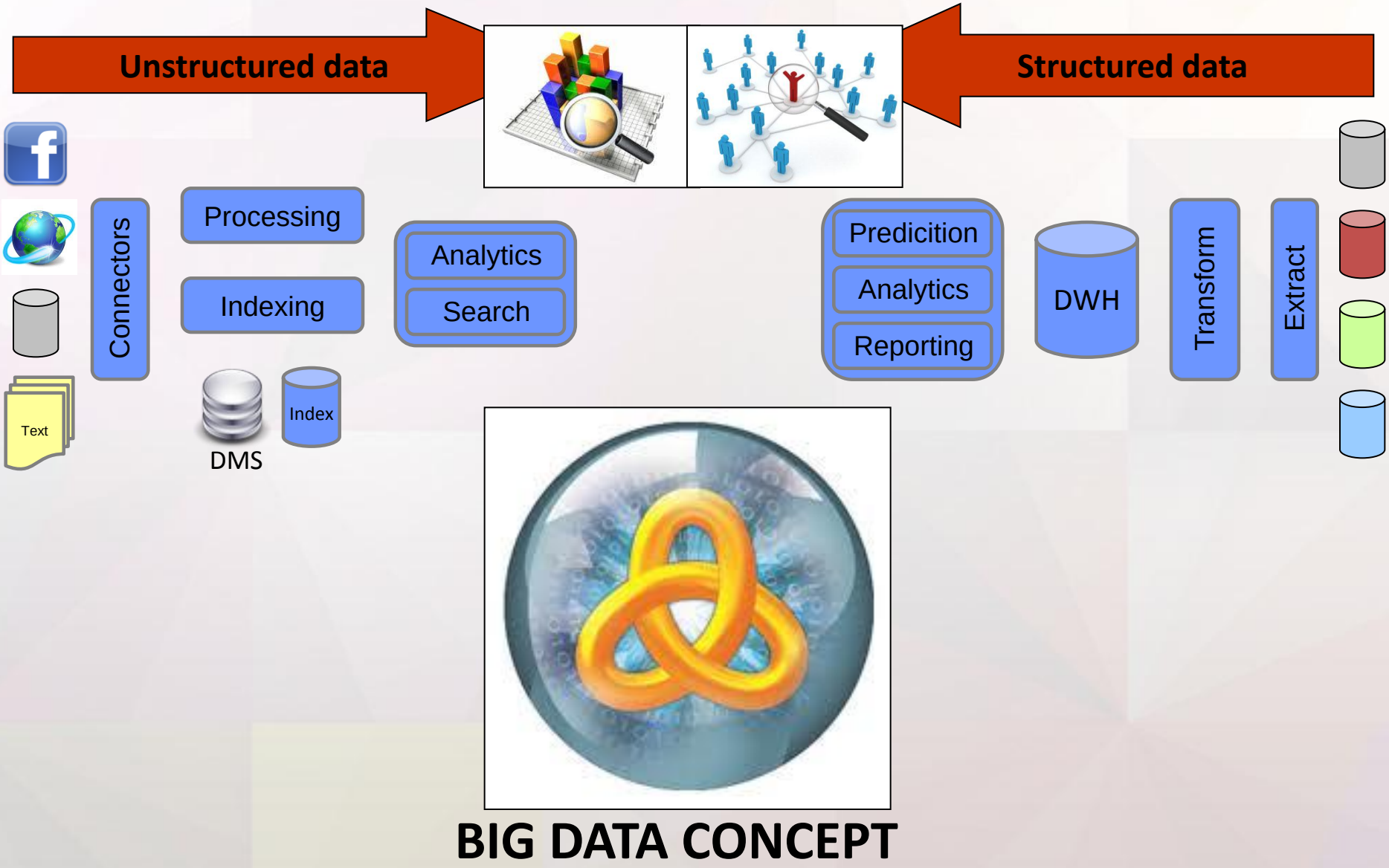
IBM

Analytical process



BIG DATA CONCEPT

Analytical process

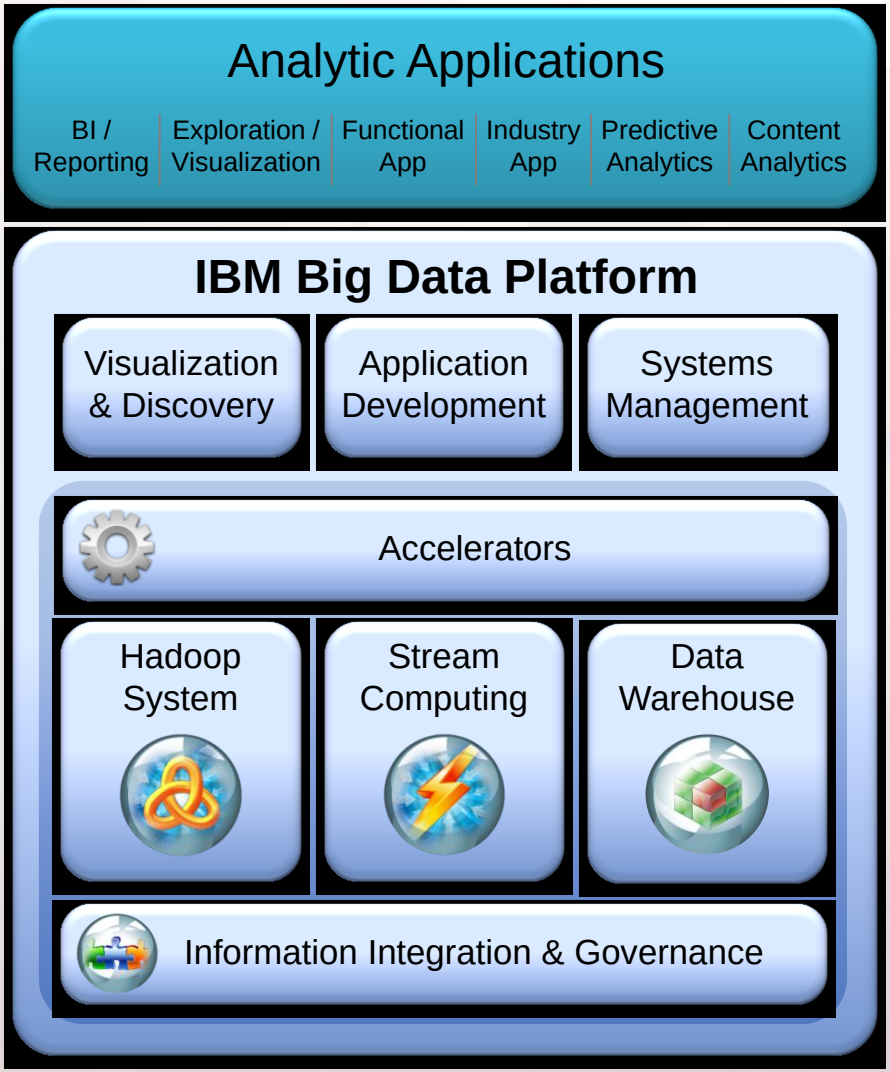


BIG DATA CONCEPT

Big Data Concept

New analytic applications drive the requirements for a big data platform

- Integrate and manage the full **V**ariety, **V**elocity, **V**olume and **V**eracity of data – **V**^{4}
- Apply advanced analytics to information in its native form
- Visualize all available data for ad-hoc analysis
- Development environment for building new analytic applications
- Workload optimization and scheduling
- Security and Governance



Hadoop

- Open source software framework from Apache
- Main concept: **BRING PROCESSING TO DATA**
- 2 Basic parts of the framework:

HDFS

- Distributed file system
- Files are split to small blocks and each block is stored on 3 places in the whole distributed system

Map/Reduce

Each task is split into 3 phases:

- Map
  - *Map* function runs in parallel on each node and returns the set of <key, value> pairs
- Shuffle
  - Pairs with the same *key* are moved close together
- Reduce
  - "Reduce" function is performed combined results for the same key together

e.g.: Count number of occuerence of words (like "IBM", "vendor", …)

- Map
  - *Map* function here is: To count number of occurences of defined words on each node (set like <"IBM", 6>, <"vendor", 8>, … is returned from each node)
- Shuffle
  - Pairs <"IBM", 6>, <"IBM, 12>, … returned from nodes are put close together for processing during the reduce phase
- Reduce
  - *Reduce* function here is summing up the final count based on partial ones returned from the nodes: => <"IBM", 6+12+…>, <"vendor", 8+9+…>

IBM BigData implemenation

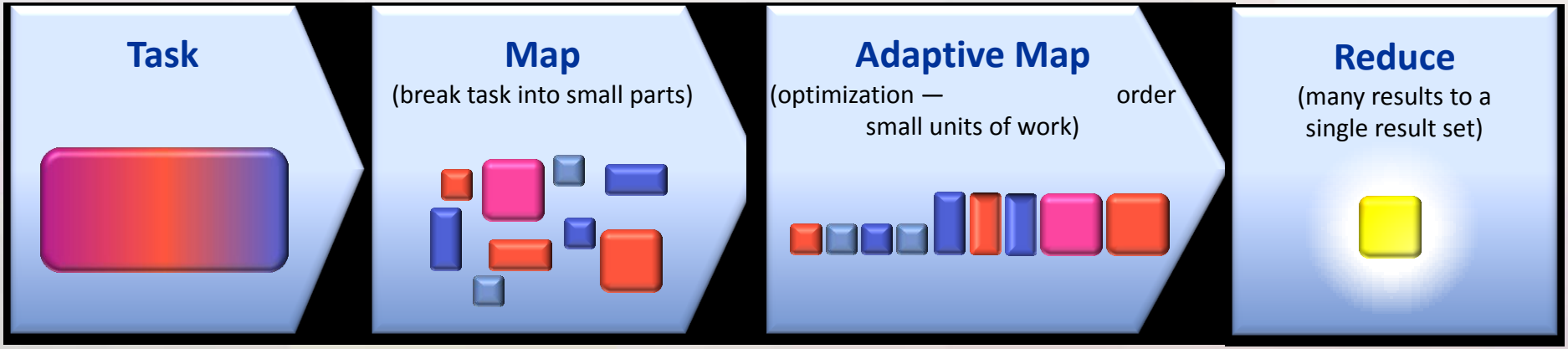
- IBM implementation of Hadoop goes much further then the classic Hadoop distributions
- First of a feature going far is **Adaptive Map/Reduce**
  - Hadoop System IBM workload optimization for hi performance

Adaptive MapReduce

- Algorithm to optimize execution time of multiple small jobs
- Performance gains of 30% reduce overhead of task startup

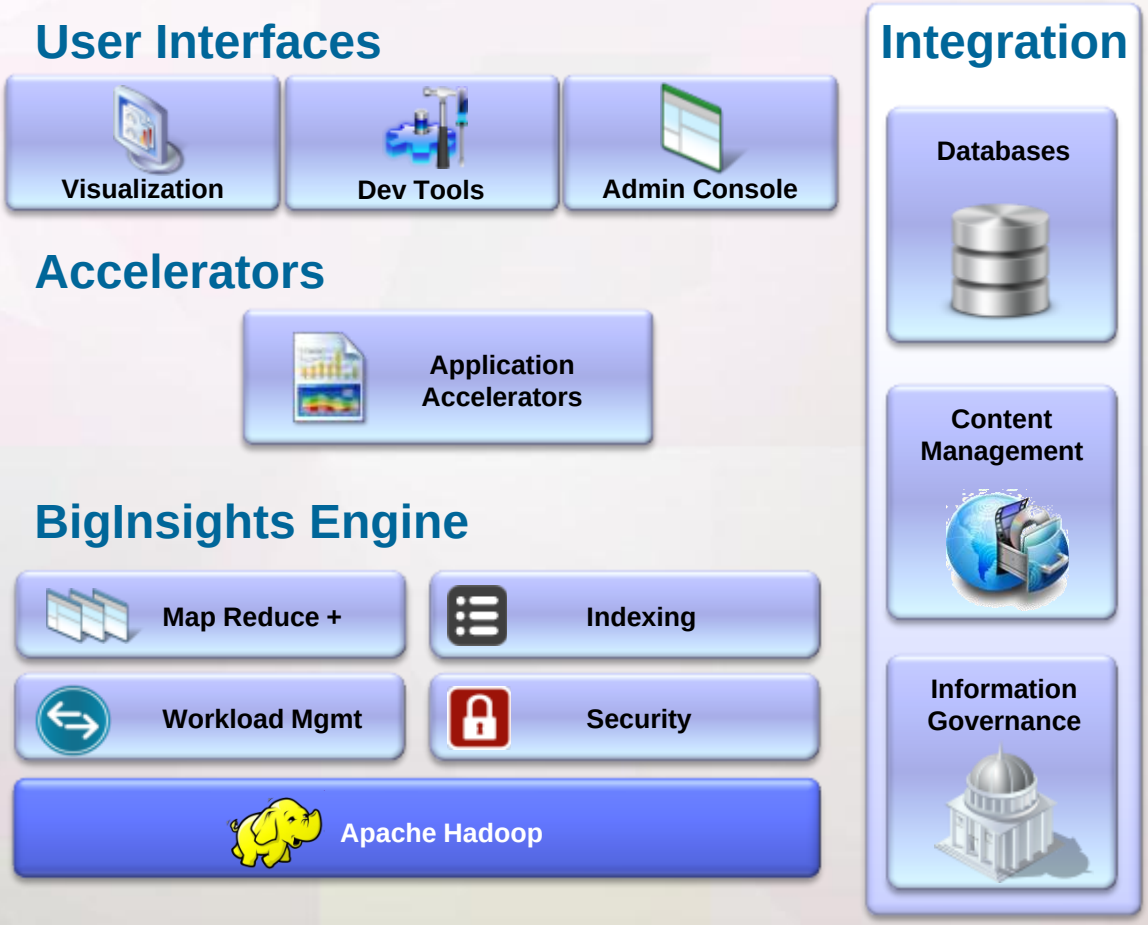
Hadoop System Scheduler

- Identifies small and large jobs from prior experience
- Sequences work to reduce overhead



IBM BigData implemenation – cont.

- Other differentiators:



More Than Hadoop

- **Performance & workload optimizations**
- **Spreadsheet-style visualization for data discovery & exploration**
- **Built-in IDE & admin consoles**
- **Enterprise-class security**
- **High-speed connectors to integration with other systems**
- **Analytical accelerators**

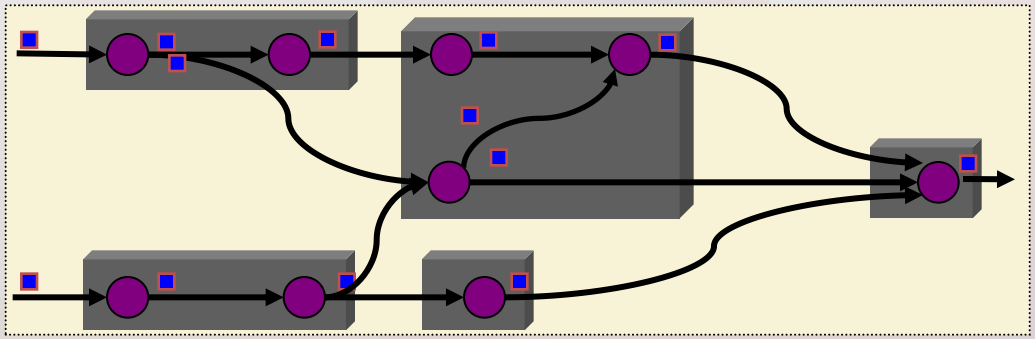
Product Name: *IBM InfoSphere BigInsights Enterprise Ed.*

Process Streaming Data

Requirement	*Technology*	*Description*
Process & Store huge volume of any data	Hadoop Map Reduce	Distributed File System *Can be used as storage and parallel runtime*
Structure and control data	Data Warehouse	Parallel Processing Engine *Can be populated by the data from analysis*
Process Streaming Data	**InfoSphere Streams**	**Stream Computing Engine** **Can be used as data source (stream of events)**
Analyze Unstructured Data	Content Analytics Text Analytics Engine	Analyze textual content for insights Used for data analysis
Integrate all data sources	ETL, Data Quality	Integrate, transform, and manage meta data Can be used for data enrichment

Process Streaming data

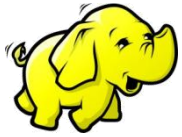
- **Technology developed with US Government**
- **Technology can execute models developed in SPSS Modeller**
- Technology is represented by *IBM InfoSphere Streams* product providing:
  - a programming model for defining data flow graphs consisting of **data sources** (inputs), **operators**, and **sinks** (outputs)
  - controls for fusing operators into processing elements (**PEs**)
  - infrastructure to support the composition of scalable **stream processing applications** from these components
  - deployment and operation of these applications across distributed **x86 processing nodes**, when scaled-up processing is required
- What's different from ETL (data pumps):
  - ETL extracts data already stored somewhere transform it and store it finally somewhere else
  - IBM InfoSphere Streams reads **big amount** of **streaming** data with **minimum latency**



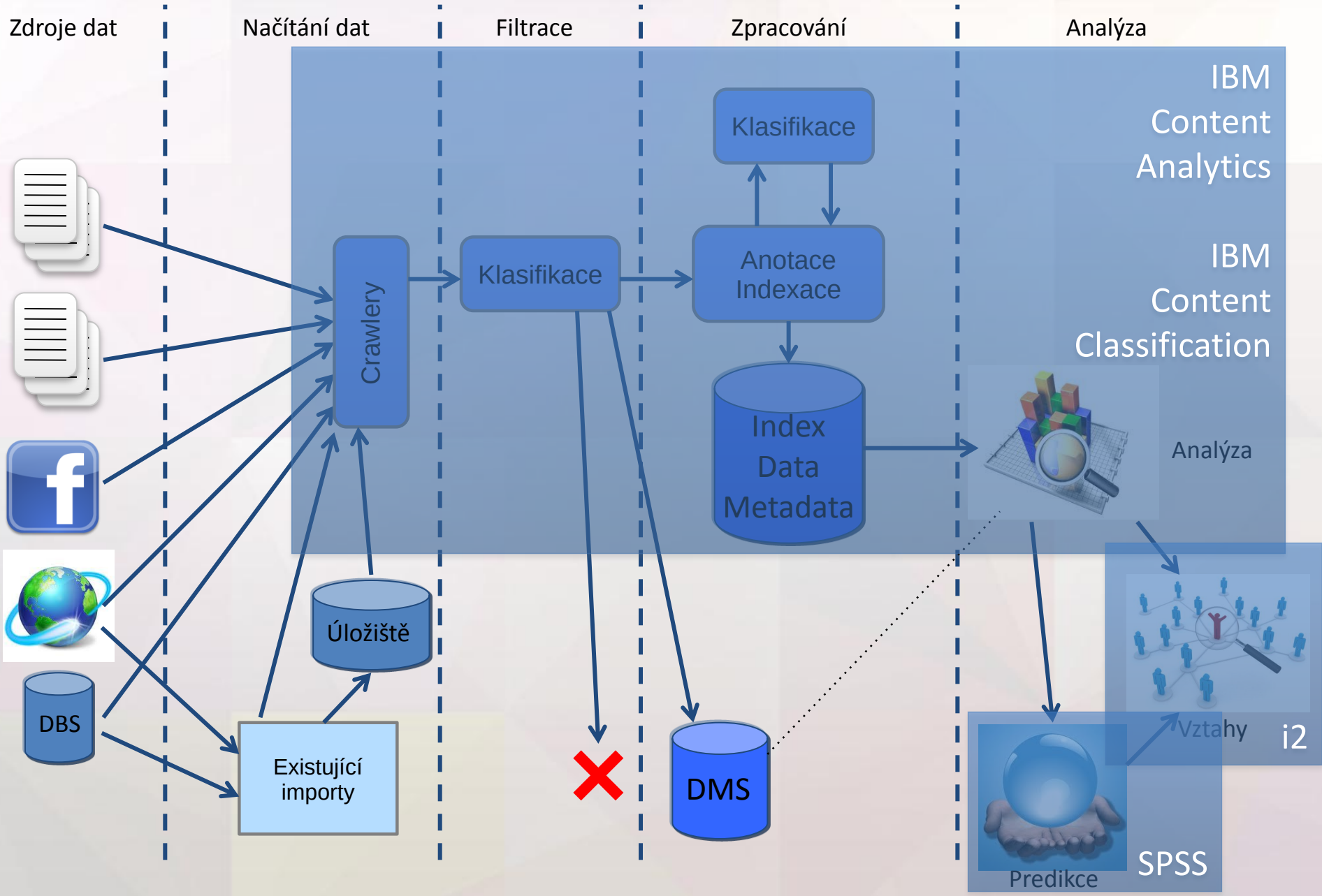
Unstructured data

BIG DATA CONCEPT

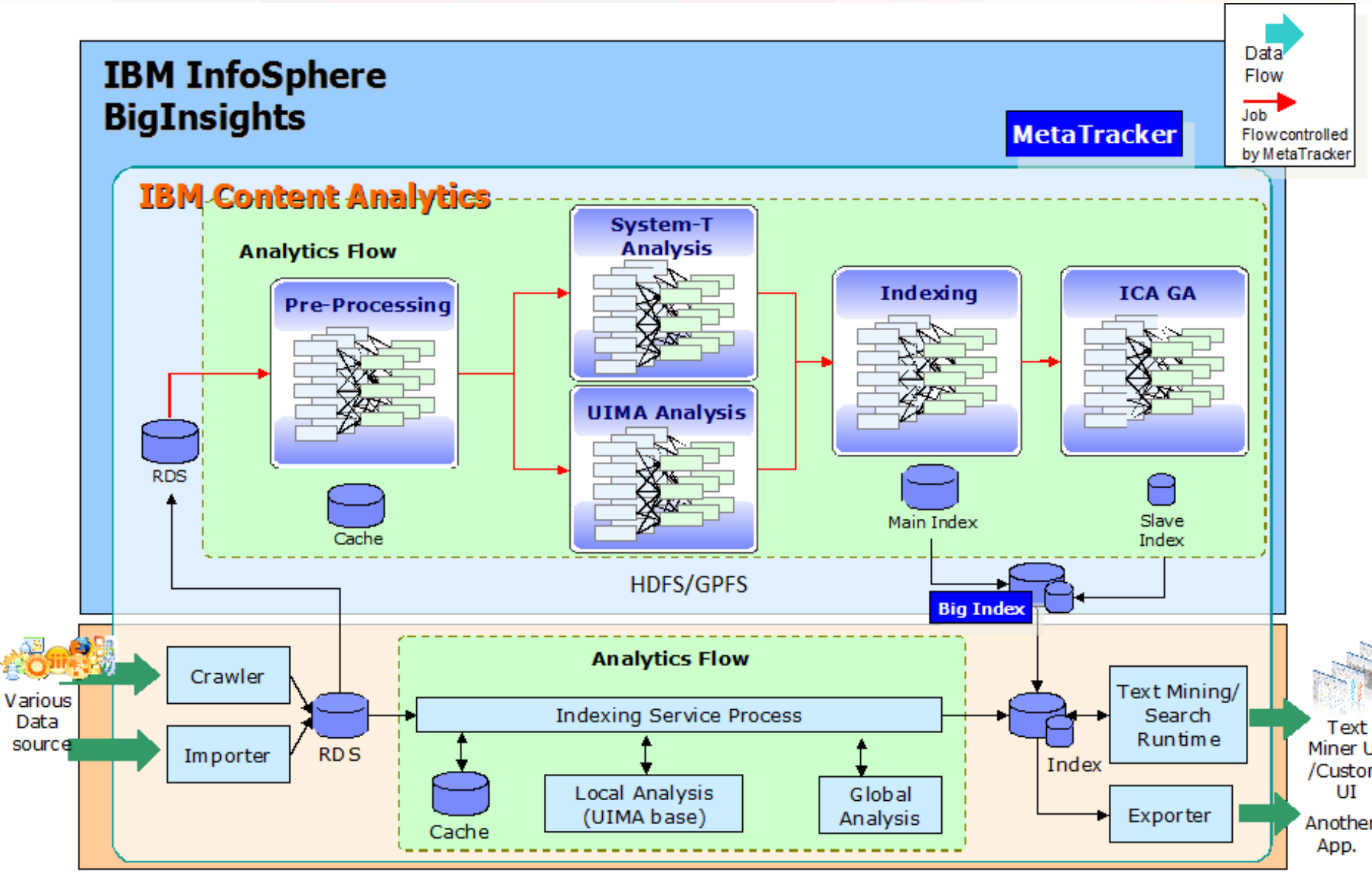
Unstructured data

Requirement		*Technology*	*Description*
Process & Store huge volume of any data		**Hadoop** **Map Reduce**	**Distributed File System** ***Can be used as storage and parallel runtime***
Structure and control data		**Data Warehouse**	**Parallel Processing Engine** ***Can be populated by the data from analysis***
Process Streaming Data		**InfoSphere Streams**	**Stream Computing Engine** **Can be used as data source (stream of events)**
Analyze Unstructured Data		**Content Analytics** **Text Analytics Engine**	**Analyze textual content for insights** **Used for data analysis**
Integrate all data sources		**ETL, Data Quality**	**Integrate, transform, and manage meta data** **Can be used for data enrichment**

Proces získání, zpracování a analýzy dat

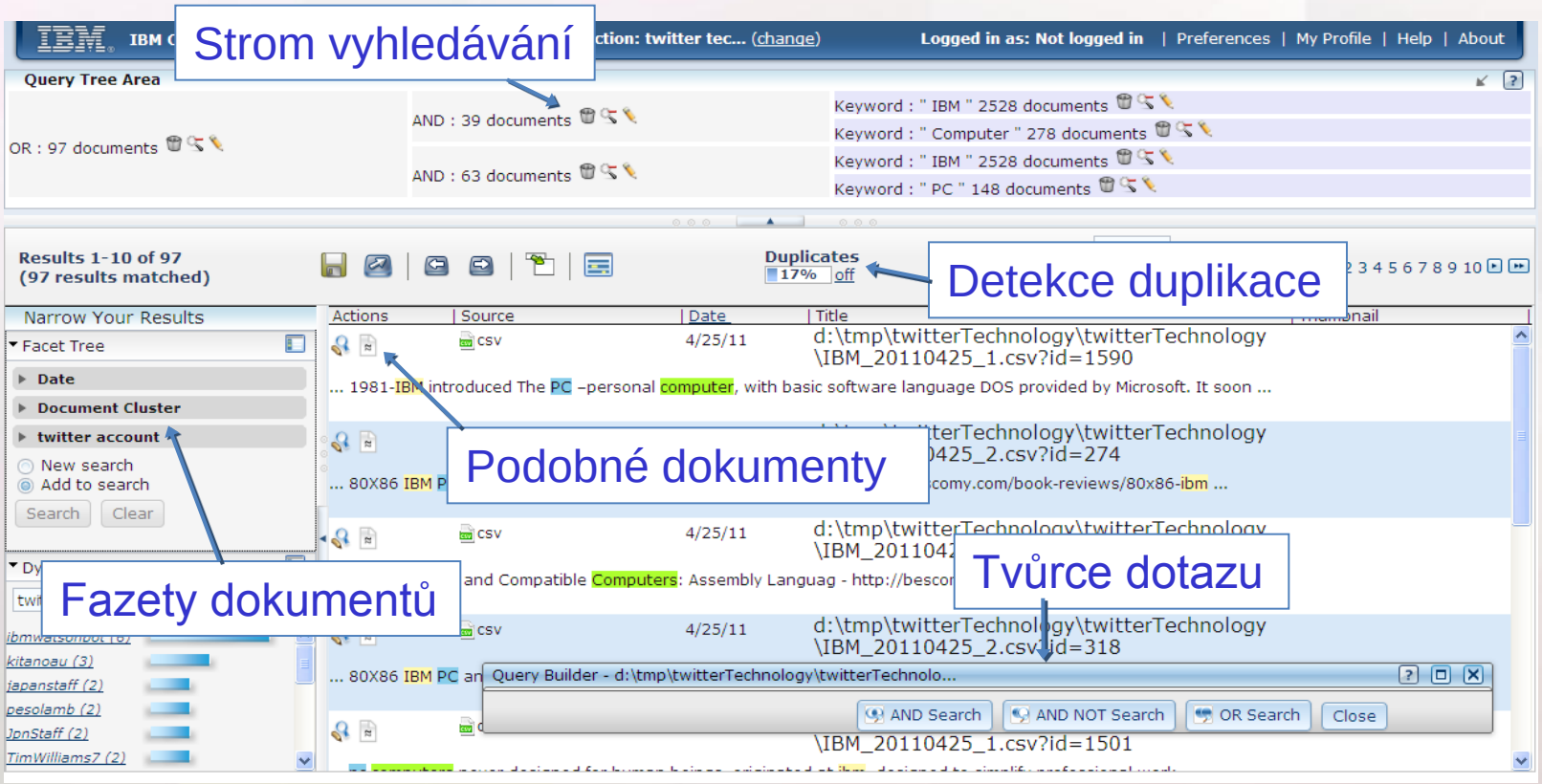


ICA BigData Support

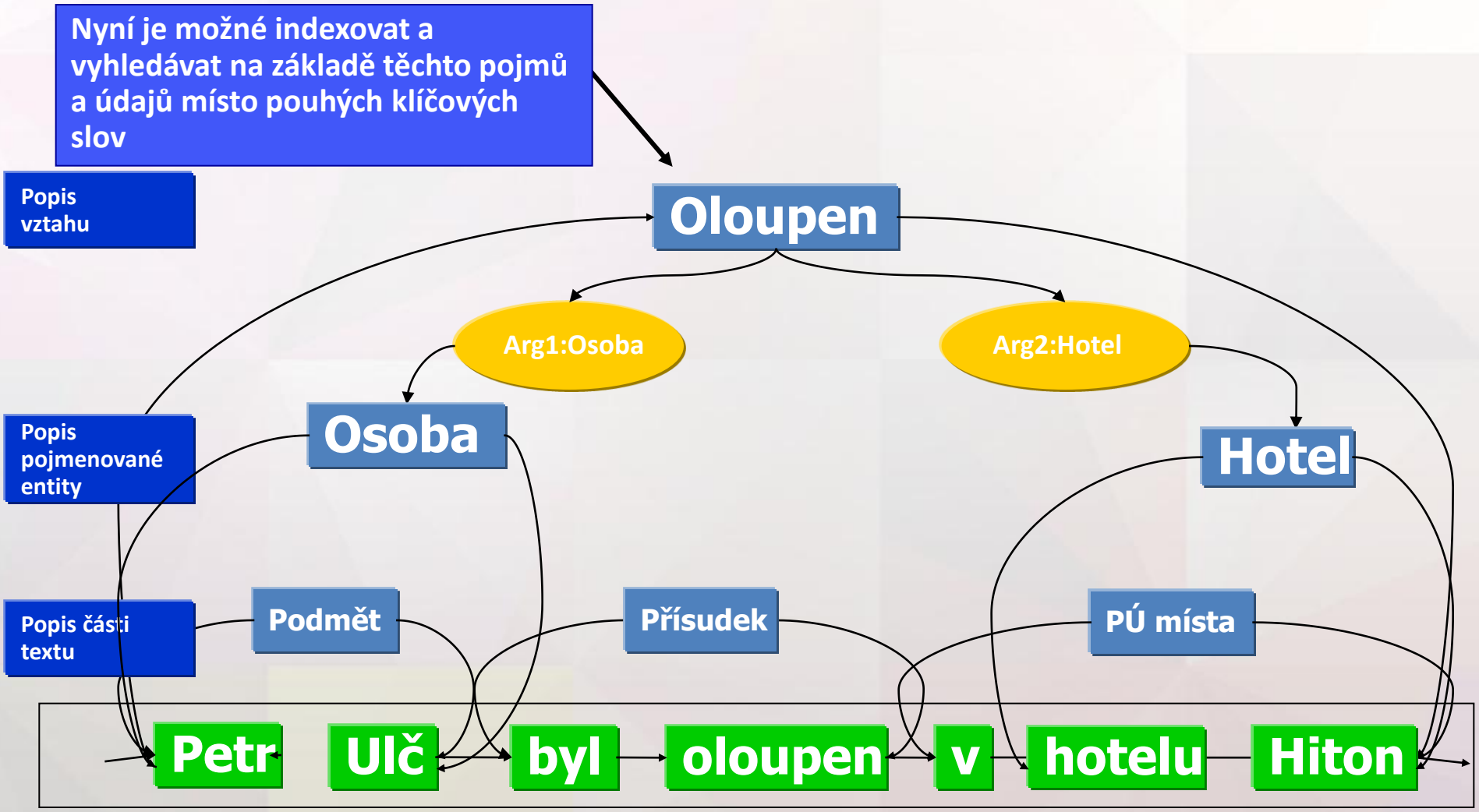


Enterprise Search

- Jednotné vyhledávání napříč organizací
  - Integrace interních i externích zdrojů vyhledávání
  - Podpora přirozeného jazyka (časování, skloňování,...)



Semantic search

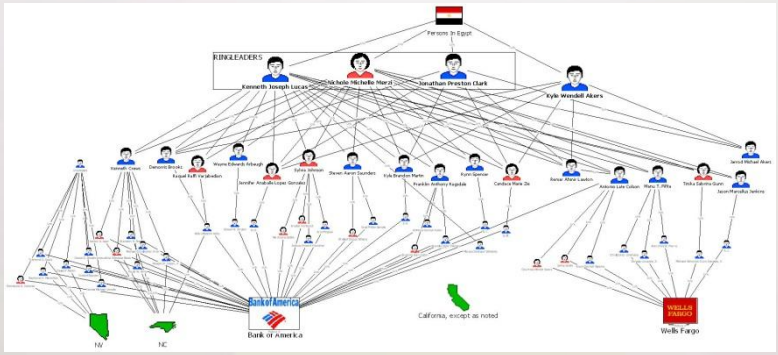


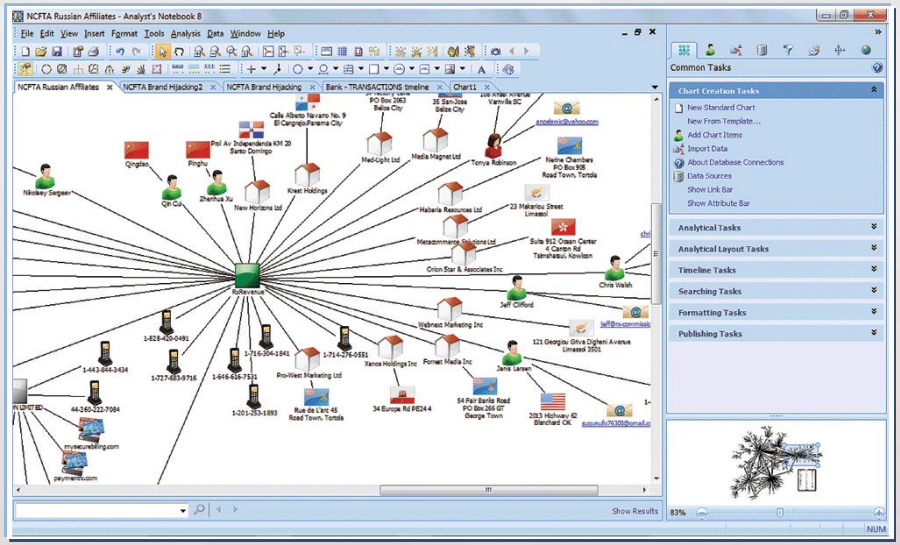
ICA – supported languages

- Arabic (ar)
- Chinese (zh)
- **Czech (cs)**
- Danish (da)
- Dutch (nl)
- English (en)
- French (fr)
- German (de)
- Hebrew (he)
- Italian (it)
- Japanese (ja)
- Polish (pl)
- Portuguese (pt)
- Russian (ru)
- Spanish (es)

Analýza vztahů - i2

- IBM i2 Intelligence Analysis Platform
  - Investigation tool
  - Data centric multi-user collaborative environment
  - Robust security architecture
  - Extensive multidimensional analysis





PROOF OF CONCEPT

IBM CONTENT ANALYTICS

Zadání POC

- Sběr dat
- Vyhledávání v datech
- Analýza dat
- Vizualizace dat, vazeb, vztahů
- Integrace, rozšiřitelnost

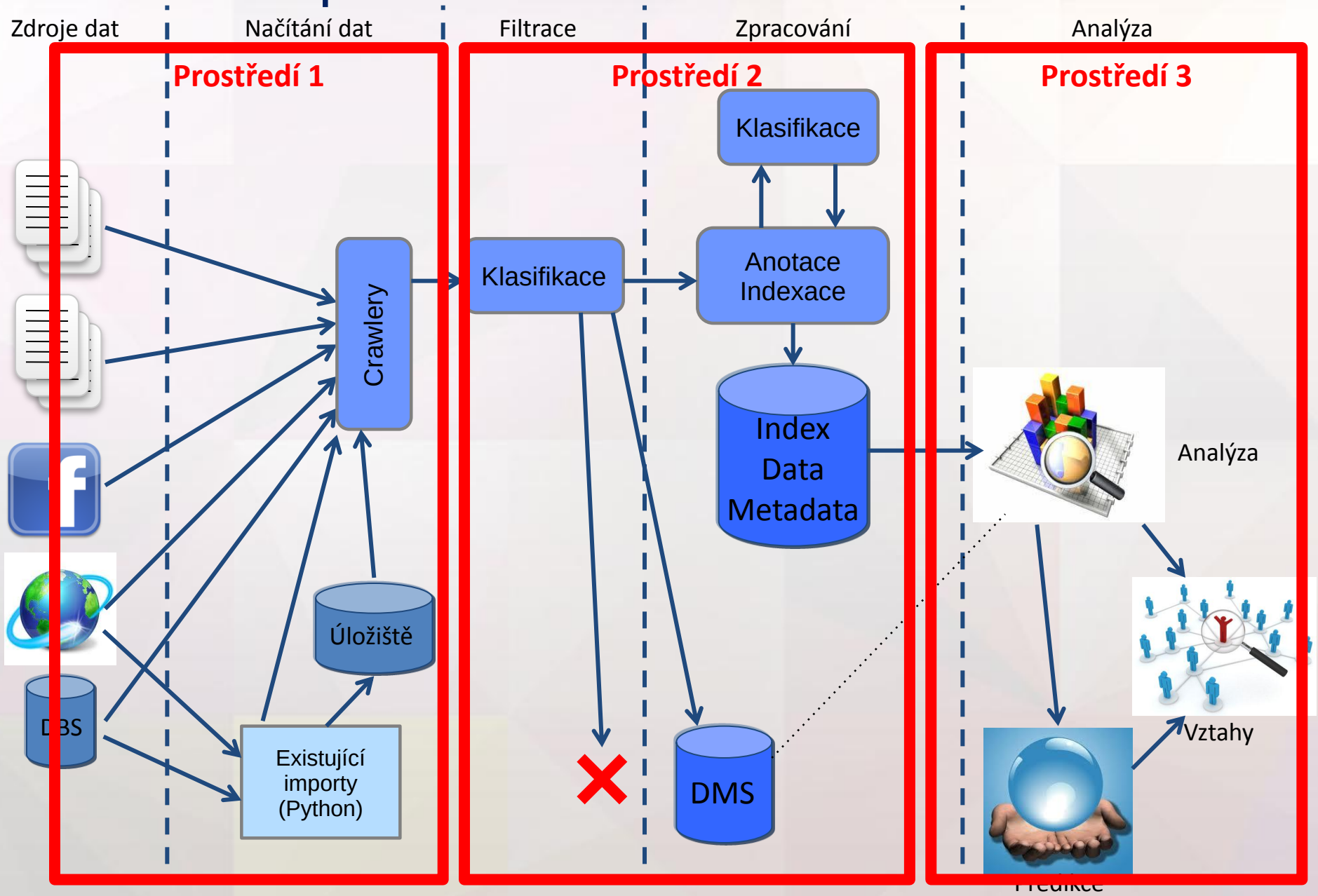
Zdroje dat

- Předané pro testovací scénáře
  - Offline soubory získané z internetu
  - Online webové servery
- Vlastní
  - Twitter
  - WAR Forum

Crawlery pro online a offline zdroje

- Webové stránky
- Soubory na disku
- Sociální sítě

Oddělená prostředí



Zpracování dat

- Unstructured Information Management Architecture – UIMA – OASIS Standard
- Tvorba slovníků
- Tvorba pravidel
- Testování

Analýza, vizualizace, integrace

- Fazety
- Časové řady
- Vazby mezi fazetami
- Duplicity
- „Značkování“
- Integrace s i2 Analyst Notebook

Vícejazyčné vyhledávání

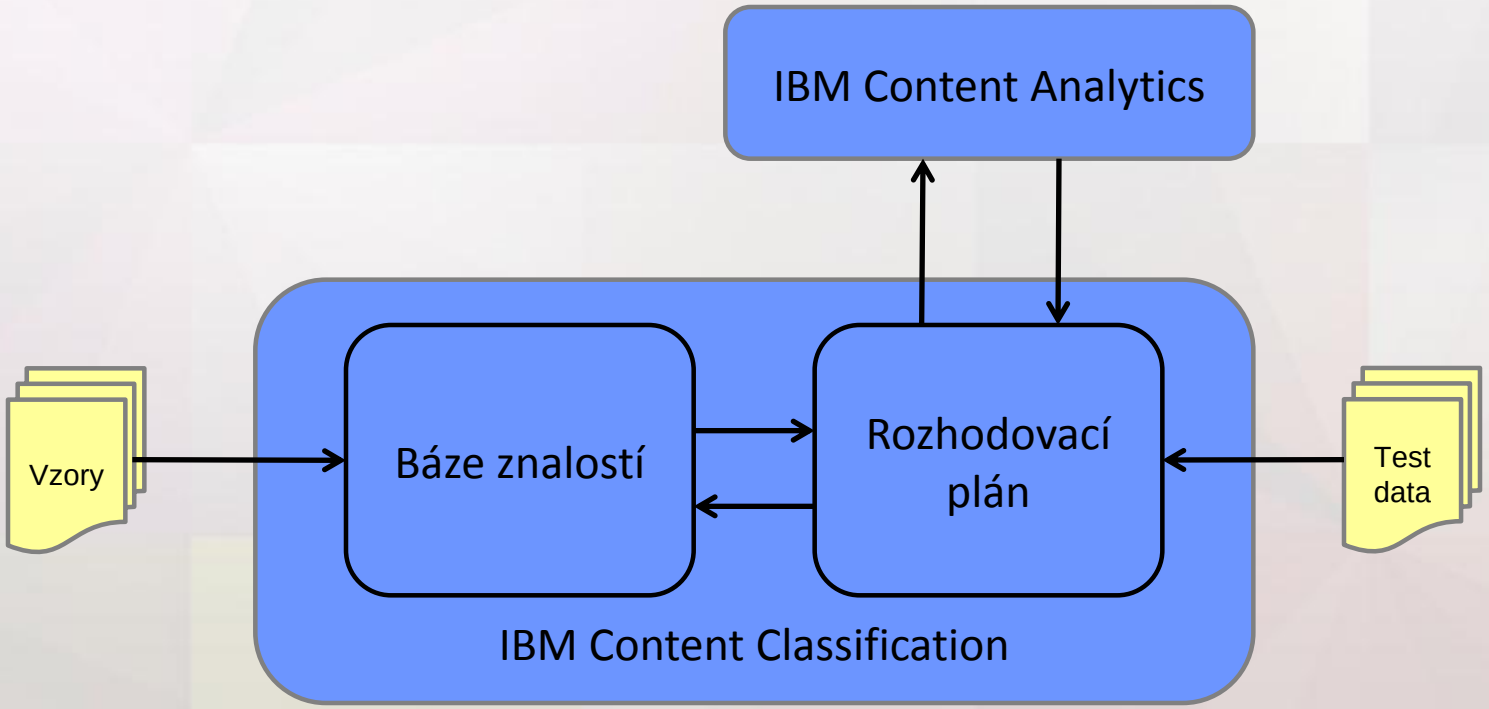
- Tvorba významových témat v ICA
- Synonymické slovníky v rámci ICA
- Externí překlad → vyhledávání v ICA
  - Offline databáze
  - Online služba

Výsledek POC

- Síla klasifikace založené na pravidlech
- Otevřená platforma včetně napojení na BigData
- Podpora českého jazyka
- Podpora platformy výrobcem v regionu

Automatická klasifikace

- IBM Content Classification
  - Učení báze znalostí se vzorových dat
  - Automatická klasaifikace obsahu
  - Adaptivní učení na základě zpětné vazby

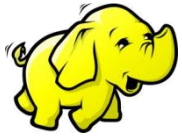


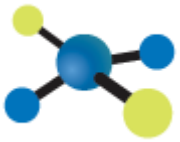
Structured data

BIG DATA CONCEPT

Structured data

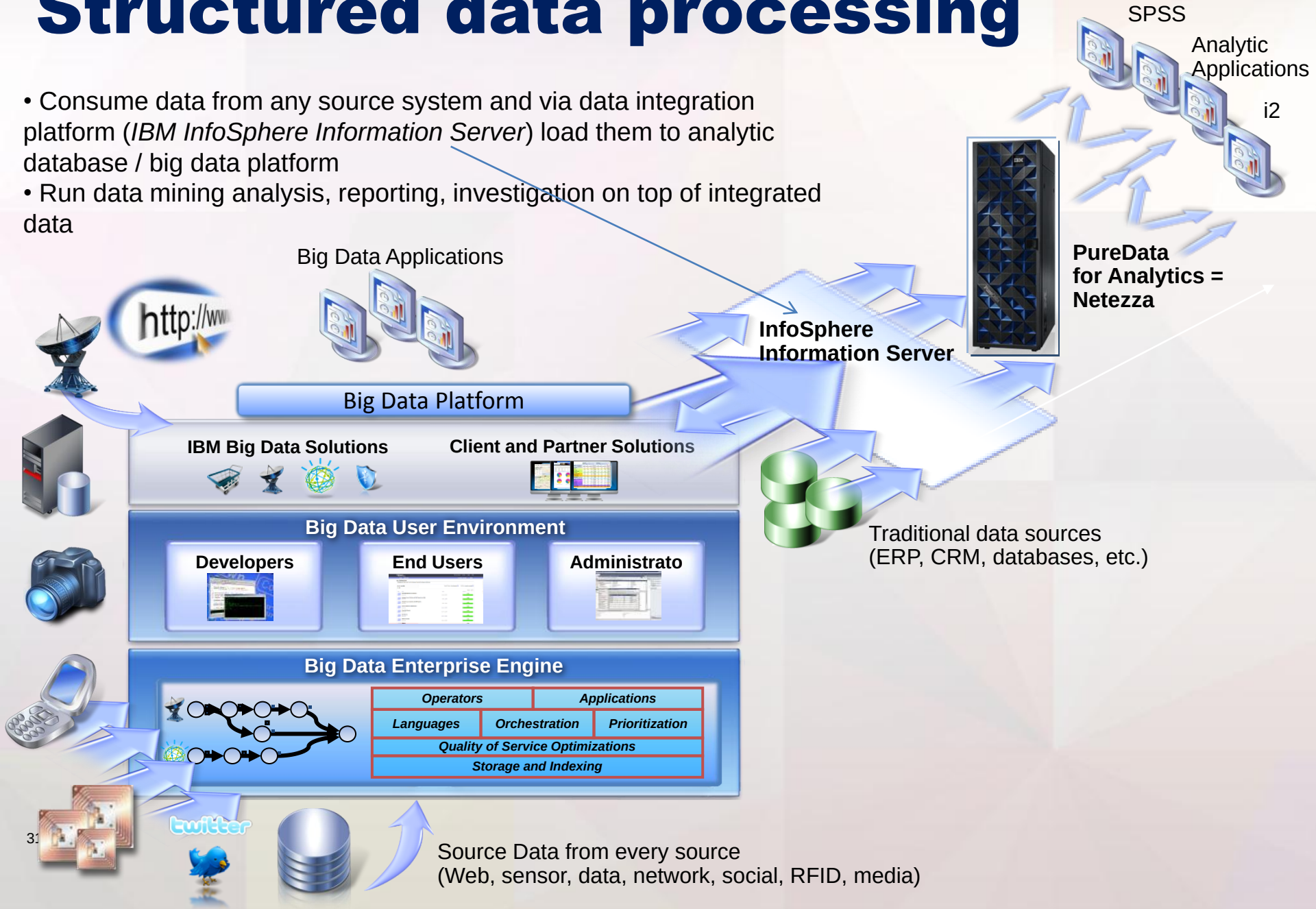
Requirement	*Technology*	*Description*
Process & Store huge volume of any data	**Hadoop** **Map Reduce**	**Distributed File System** ***Can be used as storage and parallel runtime***
Structure and control data	**Data Warehouse**	**Parallel Processing Engine** ***Can be populated by the data from analysis***
Process Streaming Data	**InfoSphere Streams**	**Stream Computing Engine** **Can be used as data source (stream of events)**
Analyze Unstructured Data	**Content Analytics** **Text Analytics Engine**	**Analyze textual content for insights** **Used for data analysis**
Integrate all data sources	**ETL, Data Quality**	**Integrate, transform, and manage meta data** **Can be used for data enrichment**





Structured data processing

- Consume data from any source system and via data integration platform (*IBM InfoSphere Information Server*) load them to analytic database / big data platform
- Run data mining analysis, reporting, investigation on top of integrated data



IBM PureData System for Analytics = Netezza

- Purpose built analytic database engine
- Appliance = HW (Server + Storage) + SW
- Very Low TCO

- Main advantages:
  - **S**peed: 10 – 100x faster then traditional systems
  - **S**implicity: **minimal administration** (no indexes, no tables spaces, …)
  - **S**calability: up to 1.2PBs for user data
  - **S**mart: **Native integration with IBM SPSS Modeller** for data mining and predictive models
    - SPSS analysis can run on the database level (no need to pass tons of data to the SPSS engine for processing)

SPSS

SPSS software and solutions enable customers to predict future events and proactively act upon that insight to drive better business outcomes

Capture

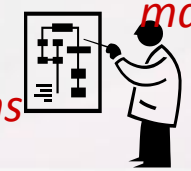
Data Collection delivers an accurate view of customer attitudes and opinions

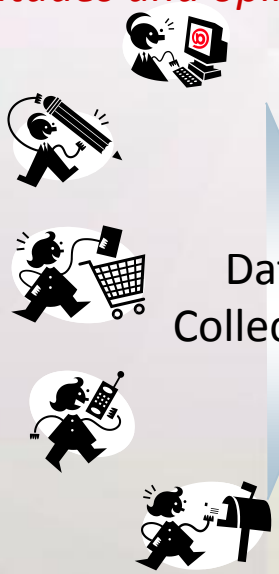
Predict

Predictive capabilities bring repeatability to ongoing decision making, and drive confidence in your results and decisions

Act

Unique deployment technologies and methodologies maximize the impact of analytics in your operation





Data Collection

Text Analytics

Data Mining

Statistics

Platform

Pre-built Content

Attract

Up-sell

Retain

Deployment Technologies

Conclusion

BIG DATA CONCEPT

SmarterAnalytics

CEE Competence Center