Text Analytics In A BigData World

BigData Florida User Group

January 11, 2016

Stanley J. Mlynarczyk, Ph.D.
Chicago Technologies Incorporated

Data Growth

--- Immense data growth requires storing of vast amounts of data

"An IDC Digital Universe study estimates amount of digital data created per year will be 35 zettabytes by 2020..."

--- Publishedby EMC Digital Universe with Research & Analysis by IDC "The Digital Universe of Opportunities: Rich Data and the Increasing Value of the Internet of Things" - April 2014

What this means for physical implementation ---

1GB = 1,000,000,000
1TB = 1,000,000,000,000
1PB = 1,000,000,000,000,000 (or 1,000 1TB disk drives)
1EB = 1,000,000,000,000,000,000 (or 1,000,000 1TB disk drives)
1ZB = 1,000,000,000,000,000,000,000 (or 1,000,000,000 1TB disk drives)

The Focus Of Industry Leaders

Google Research

http://research.google.com/pubs/NaturalLanguageProcessing.html

- ***At the syntactic level, we develop algorithms to predict part-of-speech tags for each word (e.g., noun, verb, adjective) in a given sentence as well as the various relationships between them (e.g., subject, object and other modifiers).***

- ***On the semantic side, we work on problems such as noun-phrase extraction (e.g. identifying Barack Obama, CEO in free text), tagging these noun-phrases as either person, organization, location or common noun, clustering noun-phrases that refer to the same entity both within and across documents (coreference resolution), resolving mentions of entities in free text against entities in a knowledge base, relation and knowledge extraction (e.g. is-a). While most state-of-the-art NLP algorithms attempt to solve these problems for data from a closed domain, here at Google, we solve them at web-scale...***

Text Analytics And Business

Business has a need to store, protect and utilize data

Recently mentioned by Forbes -

- ***"MongoDB, provider of a document-oriented database, has raised $80 million in Series G funding, led by a sovereign wealth fund, with participation from Goldman Sachs and from existing investors Altimeter Capital, NEA, Sequoia and funds managed by T. Rowe Price Associates, Inc, bringing total funding to $311 million."***

- ***"MarkLogic, provider of a NoSQL database, has raised $102M in Series F funding from Wellington Management Company LLP, Arrowpoint Partners, and existing investors Northgate Capital, Sequoia Capital, Tenaya Capital, and Gary Bloom, president and CEO of MarkLogic, bringing total funding to $175.6 million."***

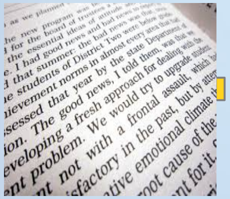
The Business Case

- The industry is moving toward deeper analytics.
- Sentiment Analysis is becoming a deeply important capability.
- The need for accuracy in text analysis in increasing
- There is a need to support low latency and very large data volumes

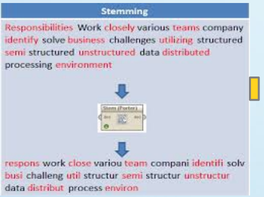
Text Analytics Basics

- Stemming – Distilling of base word forms
- Stop Lists – Ignoring of prepositions
- Word Match – Direct/exact word match
- Word Frequency – Used in document scoring

Stemming



Stop List

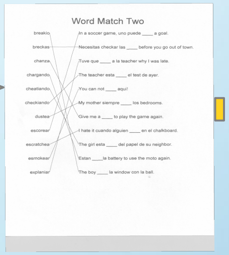


Stemming

Responsibilities Work closely various teams company identify solve business challenges utilizing structured semi structured unstructured data distributed processing environment

respons work close variou team compani identifi solv busi challeng util structur semi structur unstructur data distribut process environ

Word Match

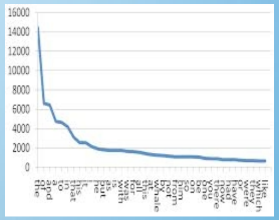


Word Match Two

breakio
breckas
chanza
chargando
cheatiando
checkiando
dustea
escorear
escratchea
esmokear
explaniar

In a soccer game, uno puede ____ a goal.
Necesitas checkar las ____ before you go out of town.
Tuve que ____ a la teacher why I was late.
The teacher esta ____ el test de ayer.
You can not ____ aqui!
My mother siempre ____ los bedrooms.
Give me a ____ to play the game again.
I hate it cuando alguien ____ en el chalkboard.
The girl esta ____ del papel de su neighbor.
Estan ____la battery to use the moto again.
The boy ____ la window con la ball.

Word Frequency



Text Analysis Challenges

→ *Accuracy/precision is challenging to implement*

→ *Accuracy/precision is challenging in real time.*

→ Search/retrieval/analysis requires massively parallel systems.

→ Existing analytics do not scale well with very large data.

→ Need for easy out-of-box text analytics for today's massive data.

→ Need to keep costs low (acquisition and operational).

The Case For Disambiguation

Examples:

“I thought your service was equivalent to getting a *root canal*.”

“The *endless stream* of *trash* on your site is *without a doubt* the most wasteful use of *tax dollars* I have ever *come across*.”

“You and your *organization* can *take a hike* the next time you are *looking for hand outs*.”

“I've had it with *poor service and lack of accountability when I call in* to *express my concerns.*

Beyond Simple Word Match

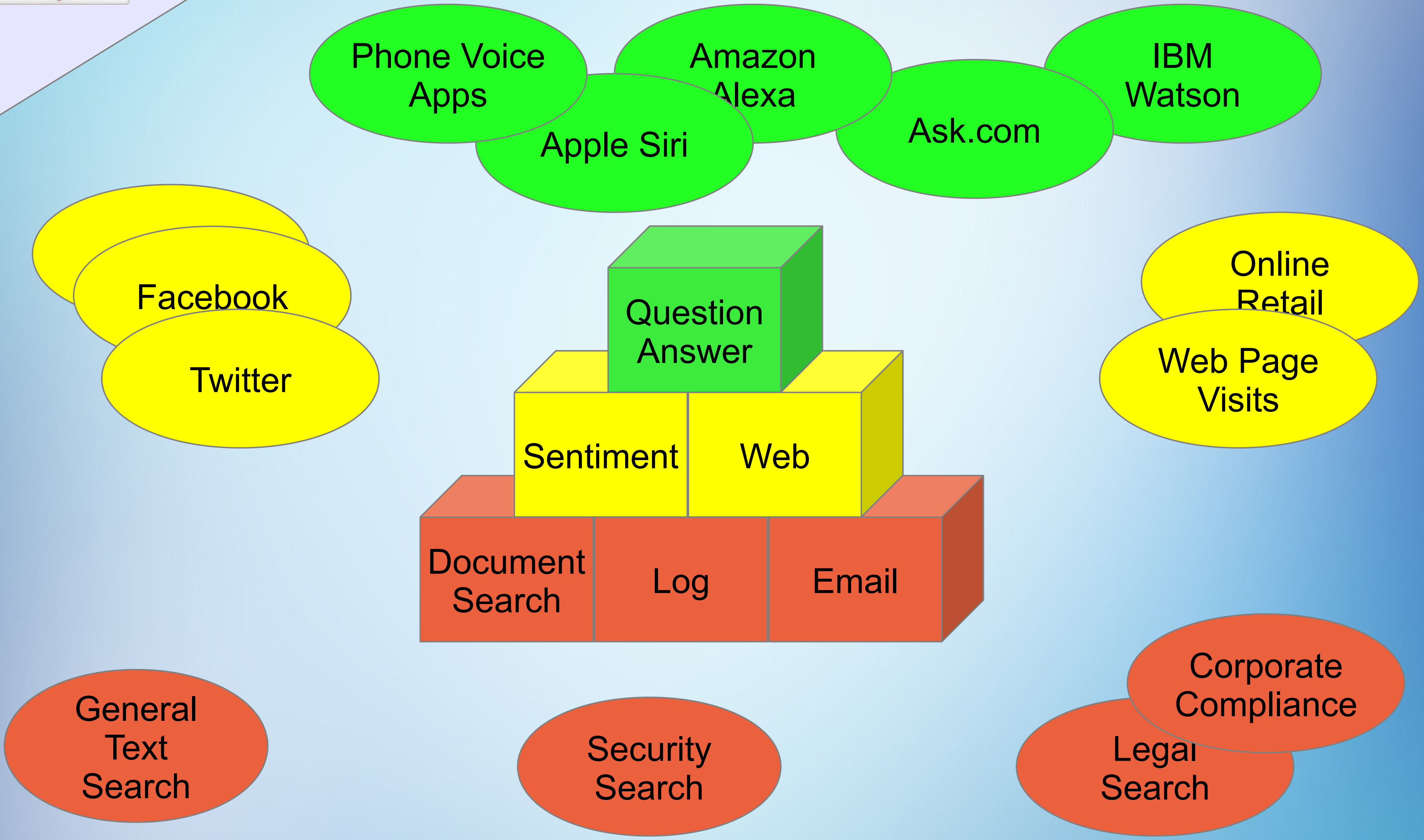
- Synonymy – Consideration of synonyms
- Part Of Speech (POS) – noun, verb, adj, etc...
- Word Sense – Resolving the true meaning of a word
- Phrase Resolution – Recognition of phrases
- Sentence Level – Grouping of words
- Sentence/Phrase resolution – Subject/Verb/Object tagging
- Fuzzy Match – recognizing misspelled words
- Inter-sentence – e.g., pronoun resolution
- Question answering – e.g., FaqFinder, IBM Watson

Considerations For Text Analytic Accuracy

- Resolve Phrases
- Resolve Synonyms
- Identify Part Of Speech (POS) – noun,verb,adj, etc...
- Identify Subject/Verb/Object
- Identify word sense within a sentence
- Identify word sense within a paragraph
- Identify named entities (dates,times,person, places, ...)
- Resolve pronouns

The computational requirements for the above are significant

Types Of Analytics



Phone Voice Apps

Amazon Alexa

Apple Siri

Ask.com

IBM Watson

Facebook

Twitter

Question Answer

Sentiment

Web

Document Search

Log

Email

Online Retail

Web Page Visits

General Text Search

Security Search

Corporate Compliance

Legal Search

Examples of Text Analytics Applications

- General document storage/search
- WEB/Cloud analytics (as a service)
- WEB analytics - sentiment, logs, ...
- Text analytics – medical, legal, call center, general
- Low latency high accuracy (semantic) text analysis
- Medical, business, government doc storage/retrieval

Email Use Case

- Businesses and government need email retention
- An ability to extract/store/catalog data
- Data protection (0 data risk) is required
- Access restriction/logging required
- Analytics/search required
- Low cost/long term storage required
- Common use case that is marketable to ALL businesses

Judicial Use Case

- All County/state/federal courts mandated to store case data
- Municipalities are drowning in data
- Massive storage requirement
- High security needed with audit controls
- No potential for data loss is a requirement
- Ability to research/analyze vast amounts of data
- Many attempts and failures (e.g. Cook County)
- Market is ripe for a solution that can perform/scale

Some Basic Tools

SPARK

SOLR

SemantiGrid

WordNet

SKB

Lucene

MongoDb

Elastic

Text Analytic Platforms

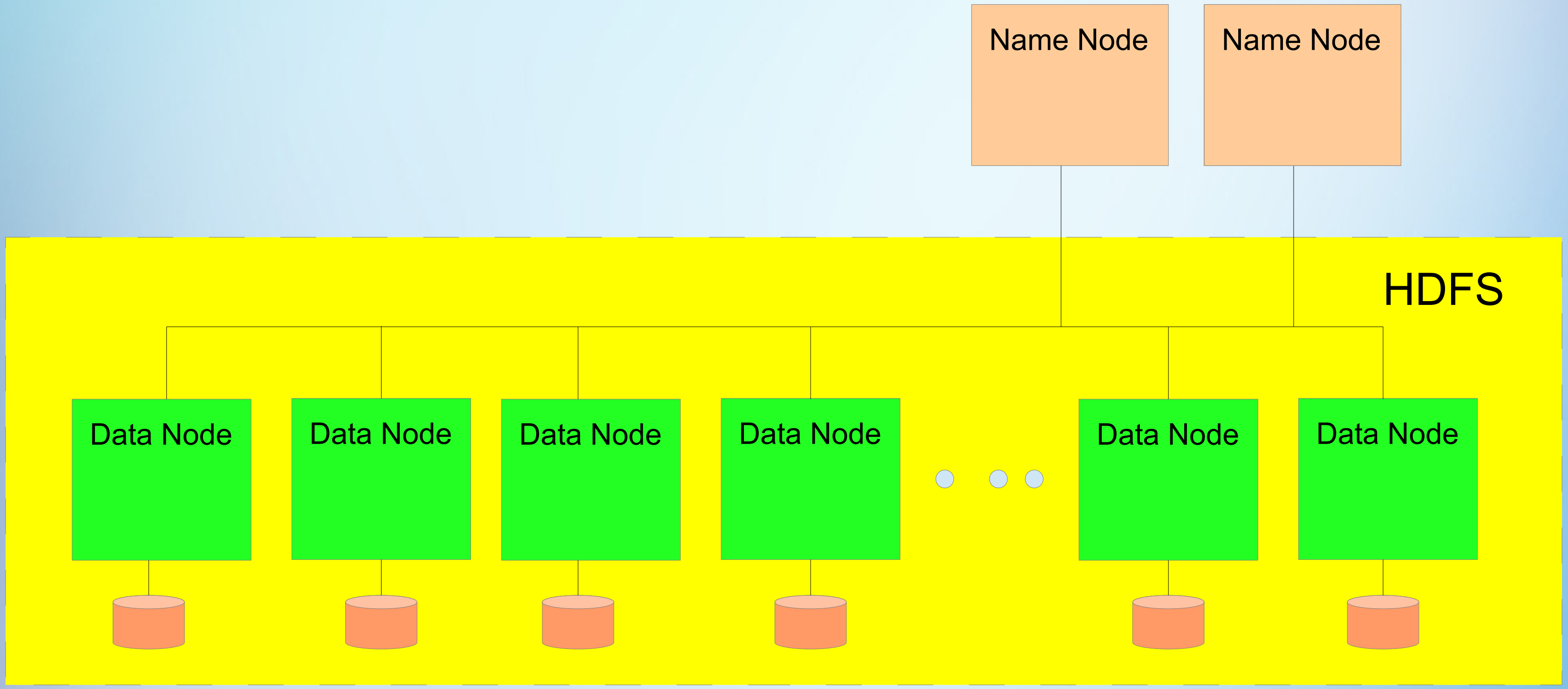
	Single System	Hadoop Cluster	Non Hadoop Cluster
Lucene	√	**2**	**2**
MongoDB	**3**	√	√
Elastic	**3**	√	√
SOLR	**3**	√	√
SemantiGrid	**3**	√	√
SGSKB	√	√	√
SPARK[1]	√	√	√

1 – SPARK included because it has some basic text/NLP capability
2 – Lucene is not parallel capable but can be used standalone on each node in a cluster
3 – These Frameworks are meant for clusters but can run standalone

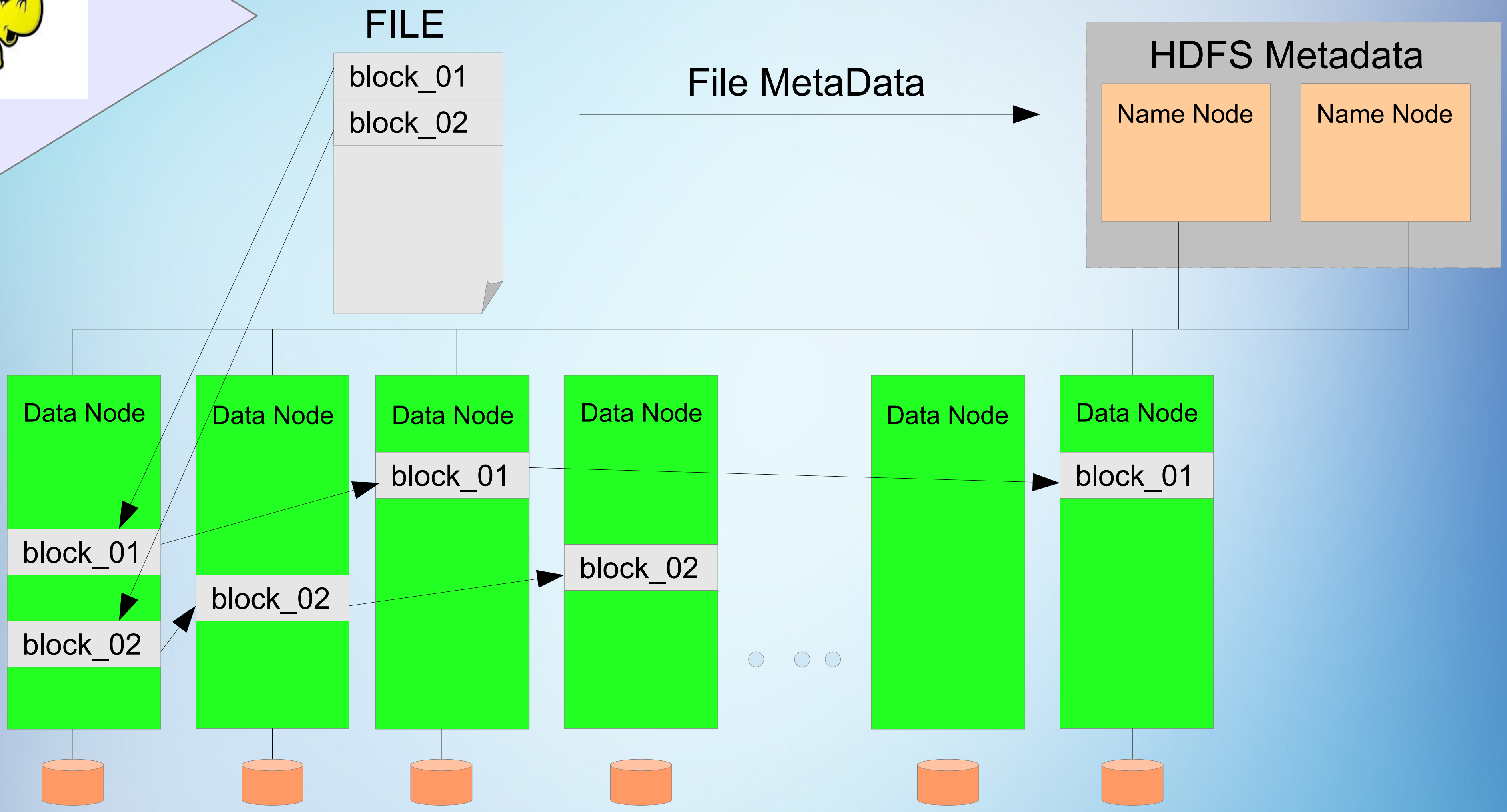
What is Hadoop?

- Compute cluster compose of data nodes and master nodes
- Distributed file system (hdfs) spanning all data nodes (Petabyte capable)
- Parallel execution capability via MapReduce and other frameworks
- Fault tolerance for files and execution
- SQL and NOSQL frameworks
- Great for multi-structured data
- Many application frameworks for AI, NLP, Machine Learning, …

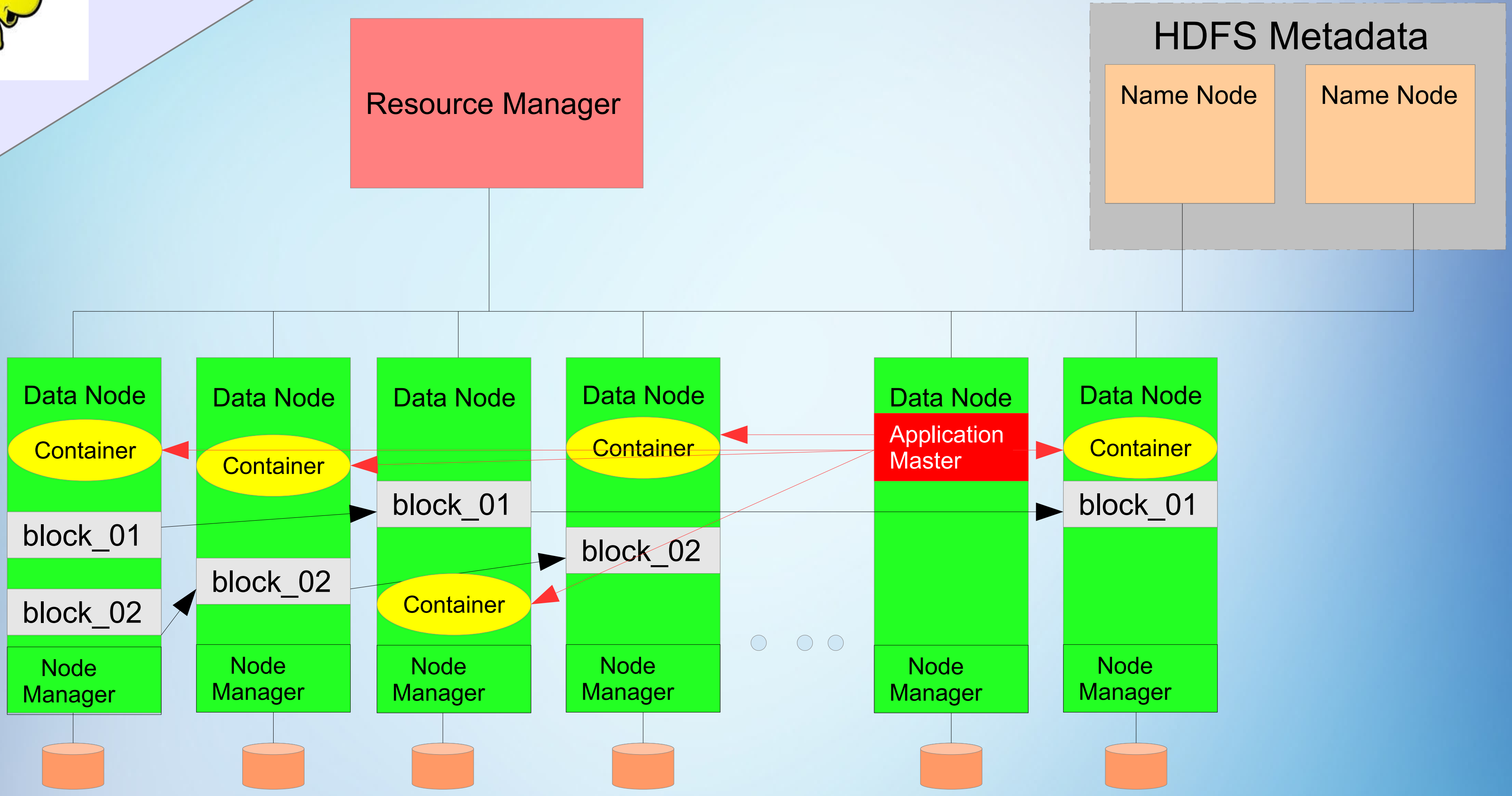
Hadoop Cluster



Hadoop Filesystem

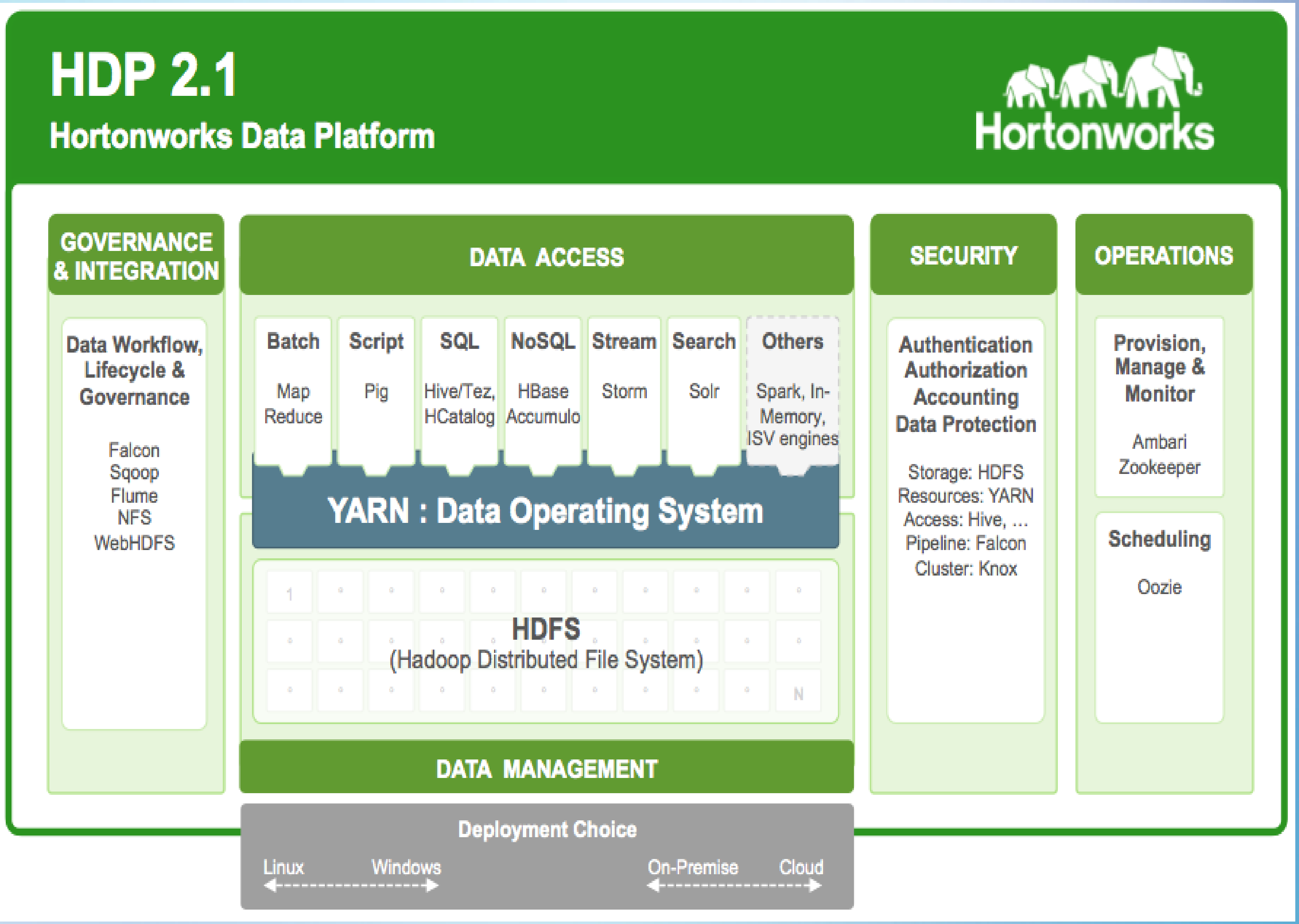


Hadoop Yarn Operating System



Resource Manager spawns an Application Master for each job which in turn negotiates for resources with Node Managers and monitors execution. Containers are individual execution threads managed by the Application Master.

Hadoop Stack



HDP 2.1

Hortonworks Data Platform

Hortonworks

GOVERNANCE & INTEGRATION

Data Workflow, Lifecycle & Governance

Falcon
Sqoop
Flume
NFS
WebHDFS

DATA ACCESS

Batch — Map Reduce

Script — Pig

SQL — Hive/Tez, HCatalog

NoSQL — HBase Accumulo

Stream — Storm

Search — Solr

Others — Spark, In-Memory, ISV engines

YARN : Data Operating System

HDFS
(Hadoop Distributed File System)

DATA MANAGEMENT

SECURITY

Authentication
Authorization
Accounting
Data Protection

Storage: HDFS
Resources: YARN
Access: Hive, …
Pipeline: Falcon
Cluster: Knox

OPERATIONS

Provision, Manage & Monitor

Ambari
Zookeeper

Scheduling

Oozie

Deployment Choice

Linux ← → Windows

On-Premise ← → Cloud

Why Hadoop For Text Analytics?

- Users may wish to store data for a variety of purposes in Hadoop
- Support multi-structured data (video, voice, structured, ...)
- Many application frameworks for AI, NLP, Machine Learning, …
- The ability to build solutions comprised of other frameworks (Pig, Hive, Spark, Hbase, ...)

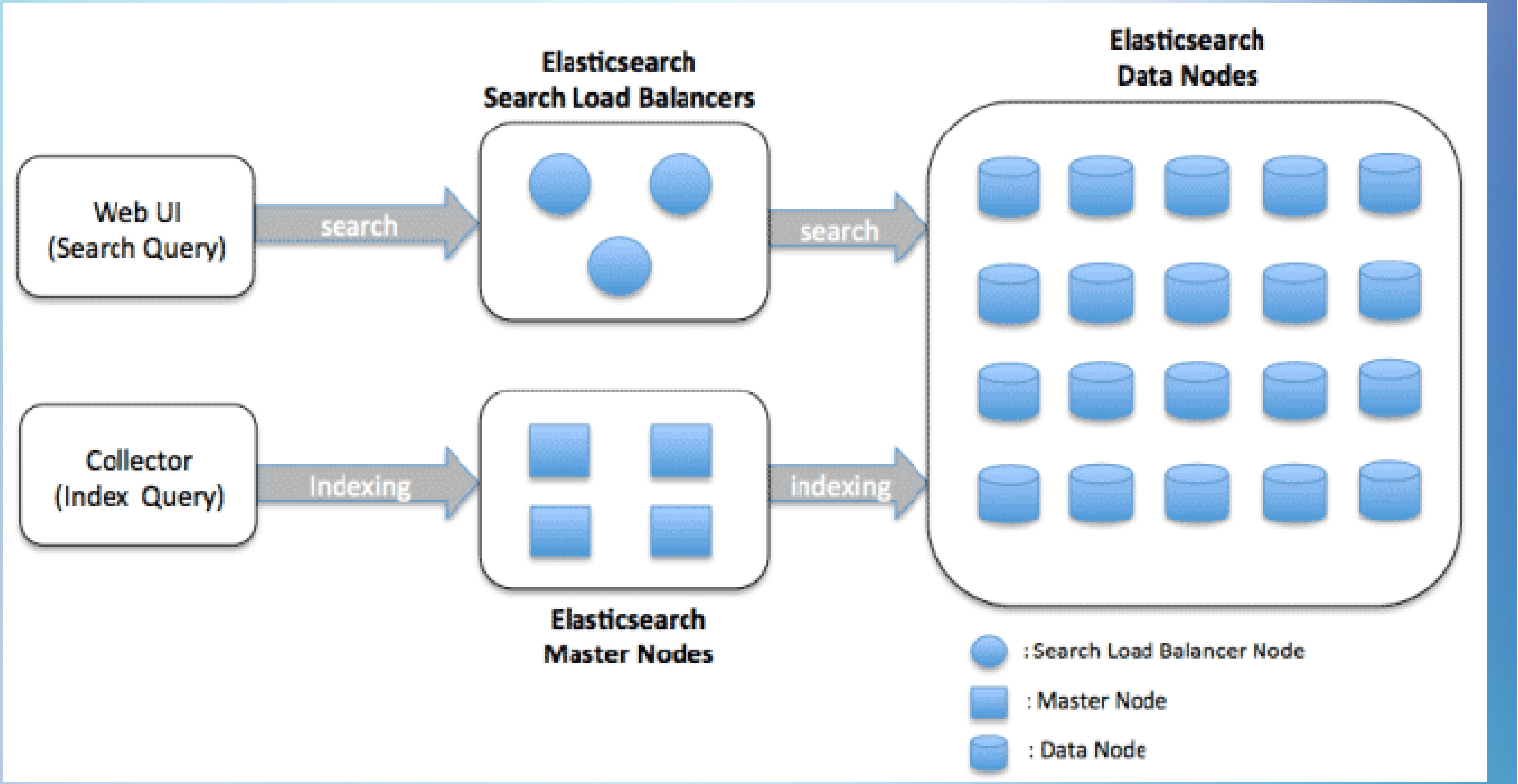
Text Analytics On Hadoop Considerations

- Performance – HDFS latency
  - Hadoop is a “big block” architecture
  - Great for scanning
  - Poor were “small block” (e.g., database functionality is required)
- SOLR, Elastic and SemantiGrid's native frameworks offer higher performance

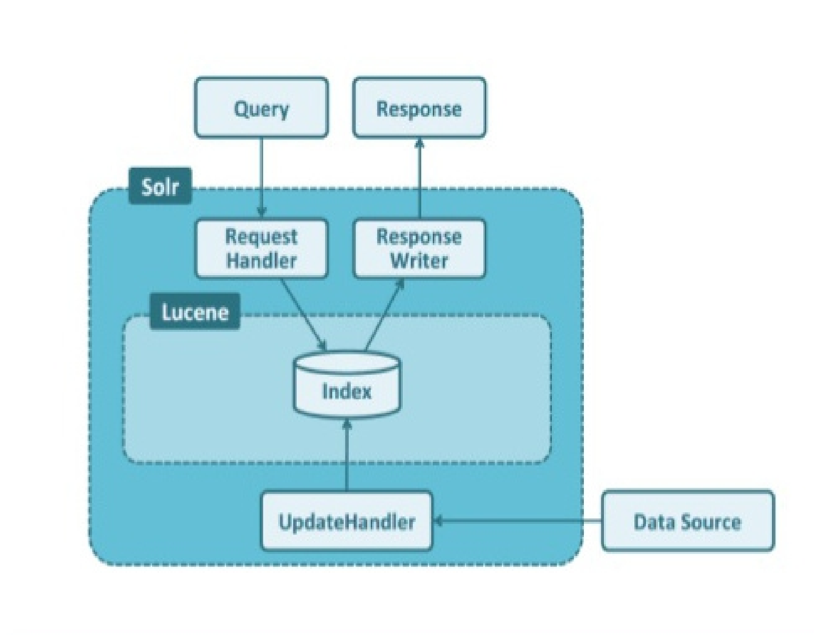
What Are SOLR and Elastic?

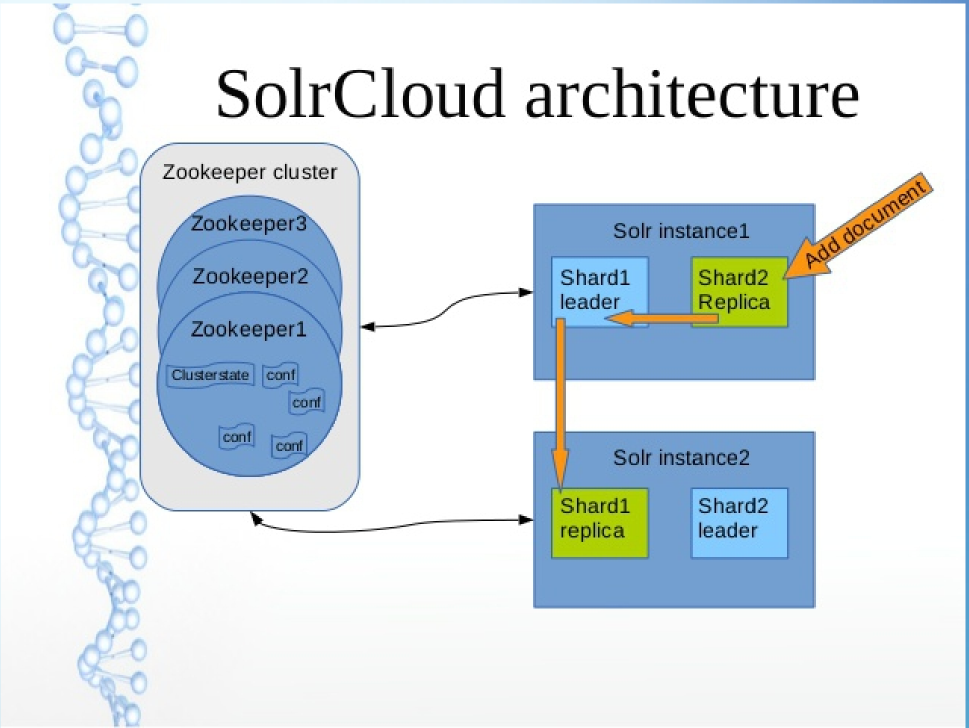
- Complete frameworks for document storage and search
- Apache projects
- Both based on Lucene
- Similar function sets
- Auto-indexing
- Full document search
- Fault tolerant
- Multi-node/scalable
- Near real-time
- Entry level semantic capability (multi-word,synonyms)

Elastic Architecture



SOLR Architecture





What is WordNet?

WordNet
A lexical database for English

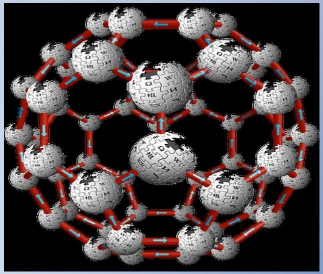
- Taxonomy based on Research of George A. Miller (Princeton)
- Support of Nouns,Verbs,Adjectives & Adverbs
- Groupings into “synsets” via synonyms
- Support for synonyms,mernonyms,antonyms,holonyms
- Support for word senses
- More ontology than taxonomy
- Comes as a callable tool + programatic API
- Open Source
- Used by researchers for advanced text/analytics & NLP
- A great way to start

What is Lucene?

- An information retrieval software library
- Apache project
- Originally developed by Doug Cutting
- Foundation for SOLR and Elastic
- Provides indexed searches of documents
- Single instance (no support for multi-node)
- Used widely within search engines

What is SemantiGrid?

(Not Yet Generally Available)



- High performance text analytics and semantic object/document store

  SKB is an integrated component

  Massively parallel/linear scaling

  Fault tolerant across geographies

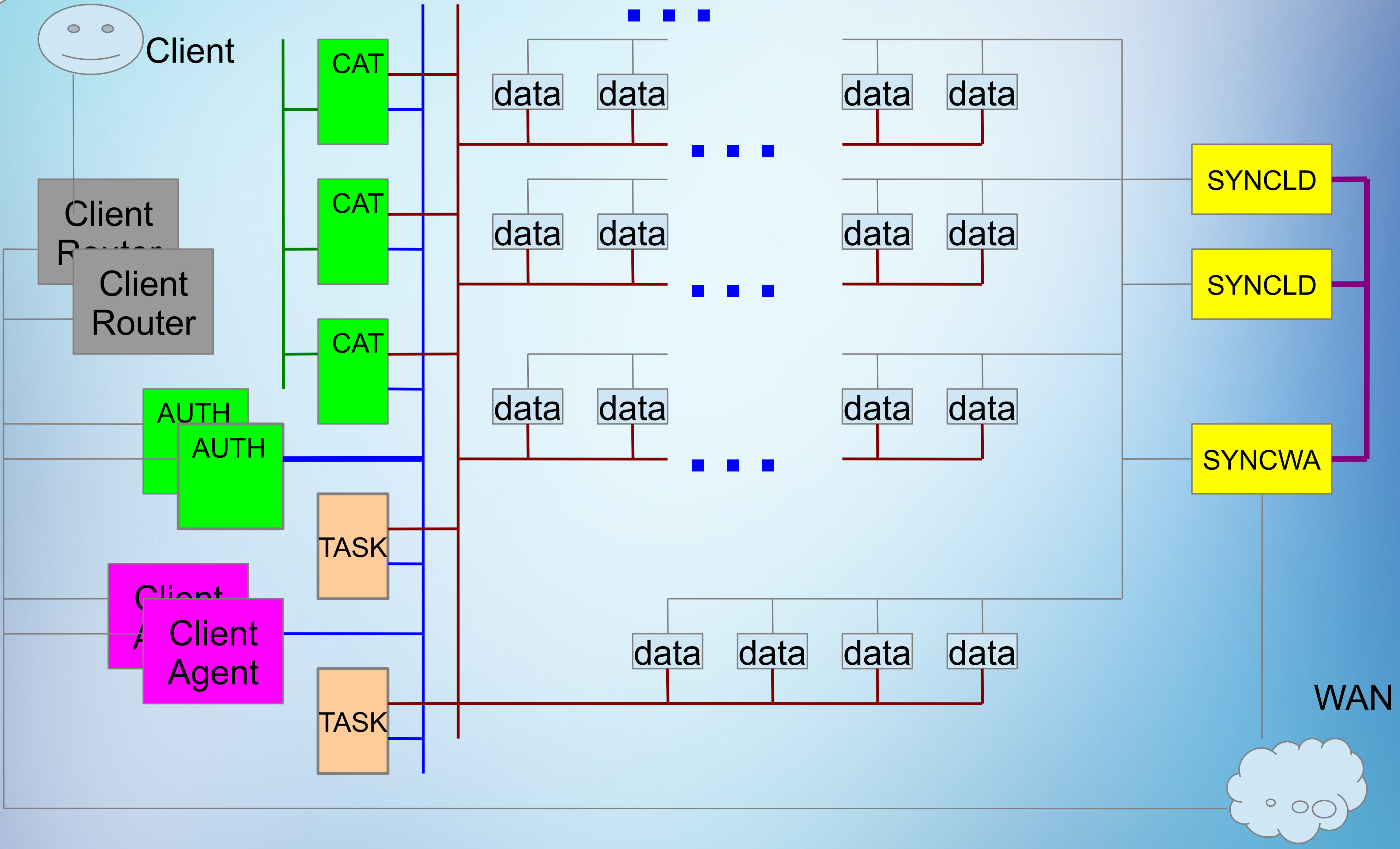
  High performance network fabric based on UDP

  Enterprise class security/audit-ability

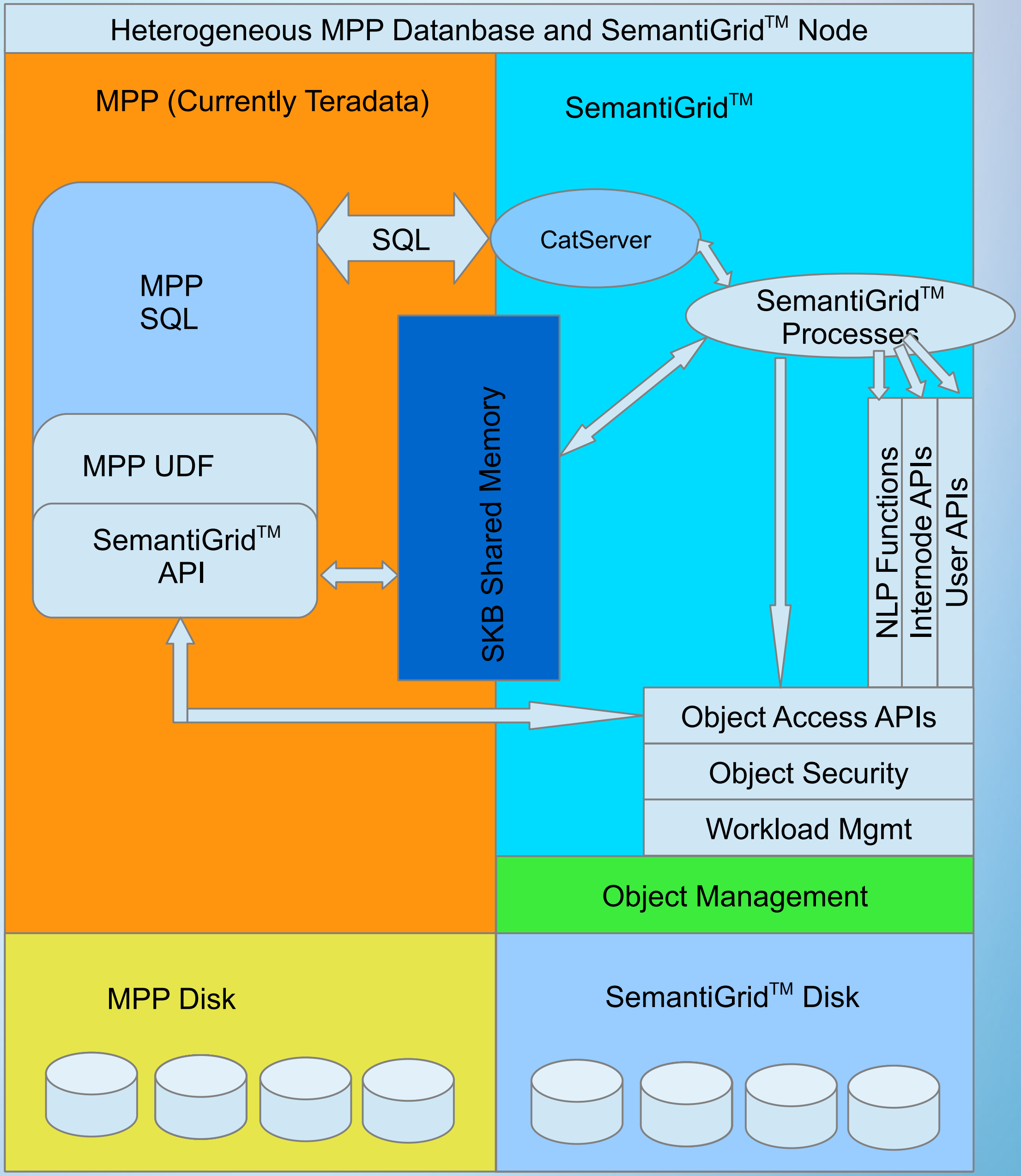
  NLP and deep semantic capabilities

  SVO and POS tagging
- Based on Ph.D. question answering research topic
- Integrates with database technology at SQL layer
- Use BI tools to access meta-data and call advanced text functions
- Usable within Hadoop

SemantiGrid Architecture



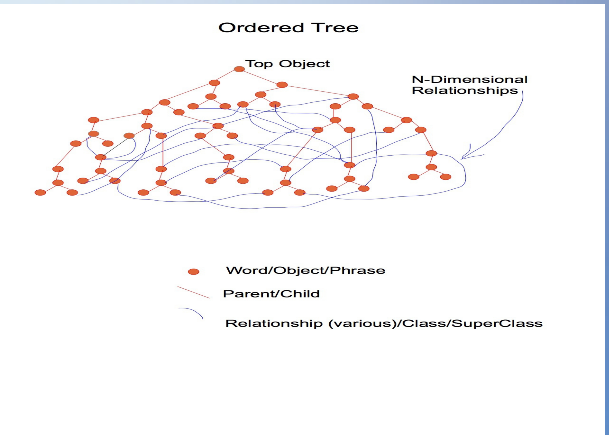
SemantiGrid MPP Integration Model*



*** Same model for most SQL databases**

What is SKB?

(SemantiGrid Knowledge Base)


Ordered Tree
Top Object
N-Dimensional Relationships
● Word/Object/Phrase
Parent/Child
Relationship (various)/Class/SuperClass

- High performance ontology

  Direct memory access

  Ordered tree

  Massively parallel/linear scaling

  N-Dimensional relationships (Graph Model)

  Available on all SemantiGrid nodes

  No operating system overhead (or context switches)

  Support for POS tagging

- Based on Ph.D. question answering research topic
- Makes semantic/syntactic analysis possible over BigData
- Usable with SQL database technology
- Usable within Hadoop
- A component of SemantiGrid

SKB Performance (nominal)

	1 CPU Core*	Single node* (8 Core)	10 node cluster * (80 cores)
Object/Word Scans	12 Million/sec	100 Million/sec	1 Billion/sec
Synonyms	938,700/sec	8 Million/sec	80 Million/sec
Word Distance	17,385/sec	140,000/sec	1.4 Million/sec

* Assumes Enterprise class Intel-Based CPUs

SKB Characteristics

- Shared memory for multi-tenancy (single persistent copy of the knowledge store)
- No operating system context switch to access SKB (***direct memory read***)
- SKB libarary functions ***do not*** require operating system calls.
- Sub-microsecond performance for many operations
- Ordered Tree structure augmented with N-dimensional relationships
- Recursive searches for high efficiency
- All data passed by reference between functions
- Currently 200,000+ word objects, 50,000 phrases + 700,000+ relationships

Tool Specifics

	Indexed Search	Semantic Capability	Streaming	Lookup Latency	Parallel Scaling	Fuzzy Match	Complete Framework	Auto Indexing	Rest API	Cluster Fault Tolerant	Domain Fault Tolerant	Data Encryption	Access Logging	Object Level Access Control	Requires Hadoop?
WordNet	√	√	X	8	X	√	X	X[3]	X[3]	X[3]	X[3]	NA	NA	NA	NO
SOLR	√	√[1]	√	5	√	√	√	√	√	√	X	X	X	X	NO
SKB (SemantiGrid Knowldge Base)	√	√	√	1	√	√	X	X[3]	X[3]	X[3]	X[3]	NA	NA	NA	NO
Lucene	√	√[1]	X	5	X	√	X	X[3]	X[3]	X[3]	X[3]	X	X	X	NO
Elastic	√	√[1]	√	5	√	√	√	√	√	√	X	X	X	X	NO
MongoDb	√	X	X	3[2]	√[2]	X	√	X	√	√	X	√	X	X	NO
SemantiGrid[4]	√	√	√	1	√	√	√	√	X	√	√	√	√	√	NO
Spark	X[3]	X[3]	√	X[3]	√	X[3]	X[3]	X[3]	√	√	X	√	X	X[3]	NO

[1] Requires Plugin [2] Doc Lookup Only [3] Not an NLP framework [4] Beta Q2/2016 NA – Not Applicable

Tool Selection Summary

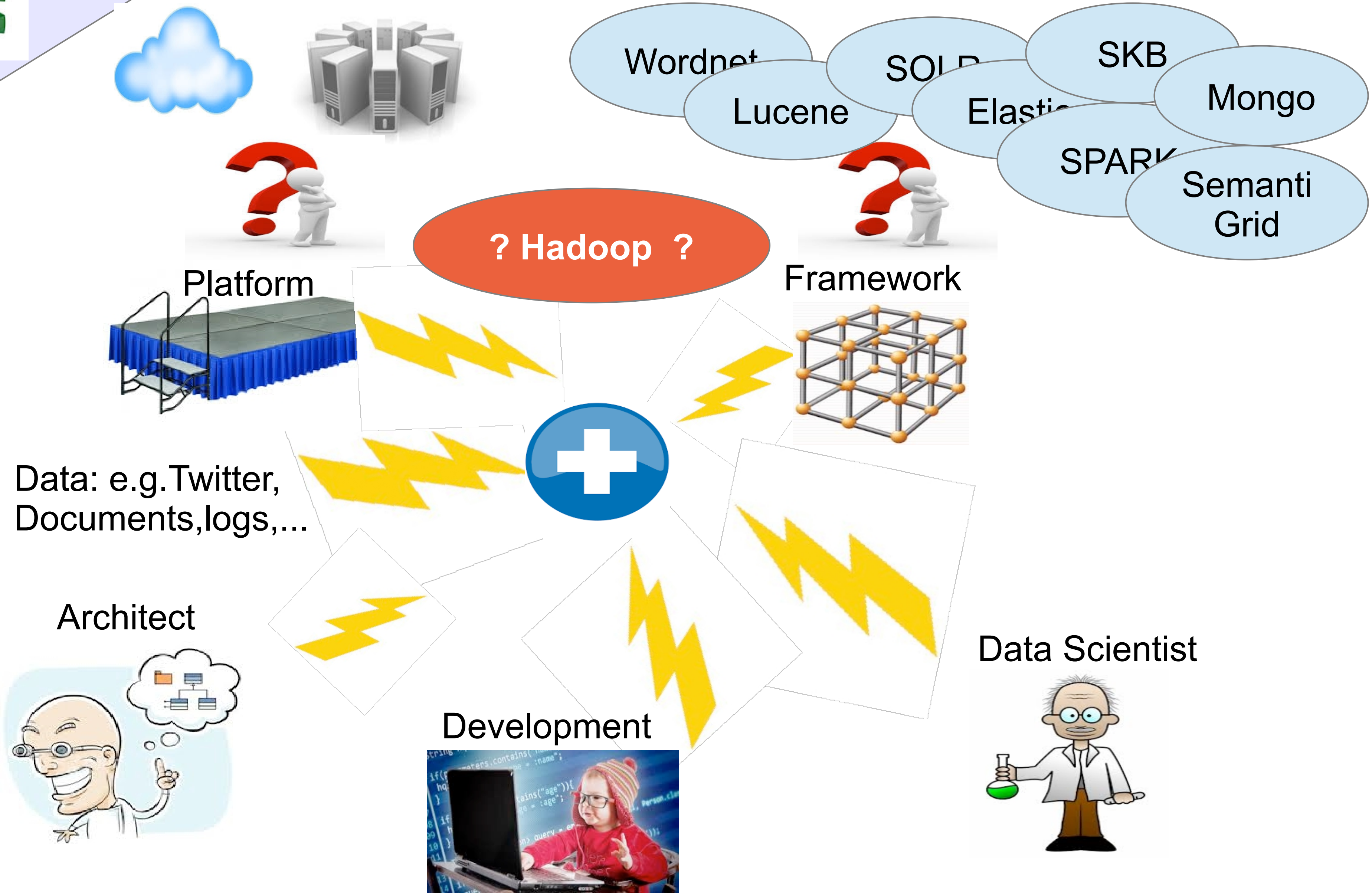
	Simple Search Small Data	Simple Search Large Data	Complex Analytics Small Data	Complex Analytics Big Data	Near Real Time Small Data	Near Real Time Big Data
WordNet	√	√[1]	√	X	√	?
SOLR	√	√	√	√	√	?
SKB	√	√[1]	√	√[1]	√	√[1]
Lucene	√	√[1]	√	√[1]	√	?
Elastic	√	√	√	√	√	?
MongoDb	√	√	X	X	√[2]	√[2]
SemantiGrid[3]	√	√	√	√	√	√
SPARK	Application Dev Platform - not an NLP framework					

[1] Local to a node

[2] Doc Ingest/Lookup Only

[3] Beta in June

Solution Components



Wordnet

Lucene

SOLR

Elastic

SKB

Mongo

SPARK

Semanti Grid

? Hadoop ?

Platform

Framework

Data: e.g.Twitter, Documents,logs,...

Architect

Development

Data Scientist

https://wordnet.princeton.edu/
https://lucene.apache.org/
http://lucene.apache.org/solr/
https://www.elastic.co/
https://www.mongodb.org/
http://www.chitechcorp.com/
http://spark.apache.org/
https://dev.twitter.com/rest/public

Stanley J. Mlynarczyk – Ph.D.
CTO Chicago Technologies Incorporated
stan@chitechcorp.com