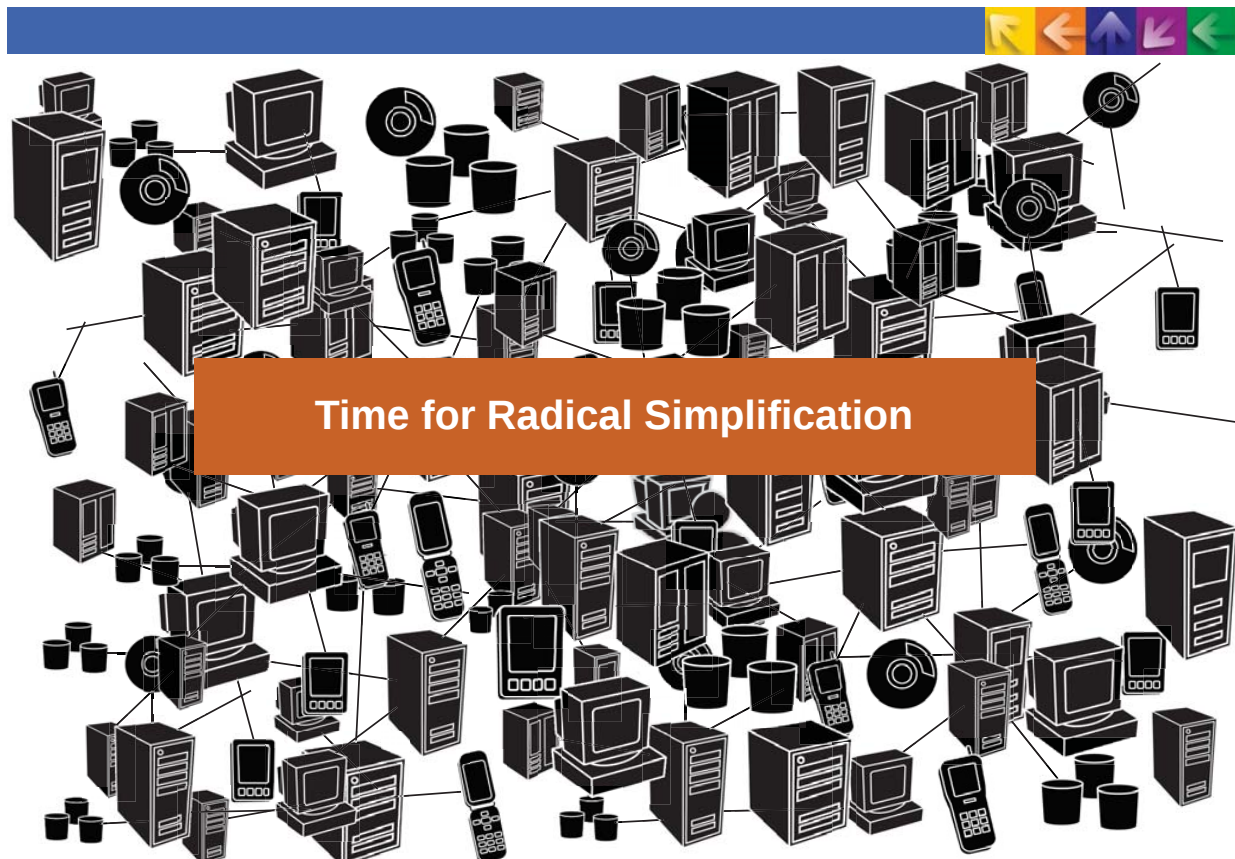


Cloud Computing and Data Quality Services



© 2010 IBM Corporation



Outline

- Cloud Computing and Hadoop Architecture
 - Cloud Applications
 - Service Models
 - Hadoop and MapReduce Architecture
 - Data Flow in Hadoop
- Data Cleansing Dimensions
- Enhancing Data Quality with Web Data

Trend: New “Big Data” becoming commonplace



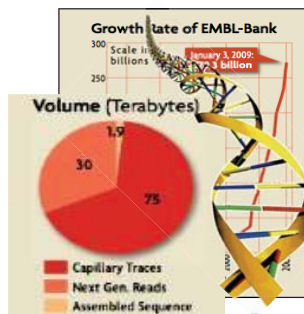
Transactions: 46 Terabytes per year

bharti

Call Records: 3 Terabytes per day



100 Terabytes per year



Genomes: Petabytes per year



LHC: 40 Terabytes per second

facebook

10 Terabytes per day

twitter

7 Terabytes per day

Google

20 Petabytes per day

You Tube

New Video Uploads:
4.5 Terabytes per day

Massive Volumes of Data at Rest and in Motion.....



Applications

- Data Integration Services
- Product Data Enrichment
- Insurance: Pay as you drive
- Currency Life Cycle
- Product Verification Services
- Address Completion and Validation
- Bringing Social Network Data in DWH

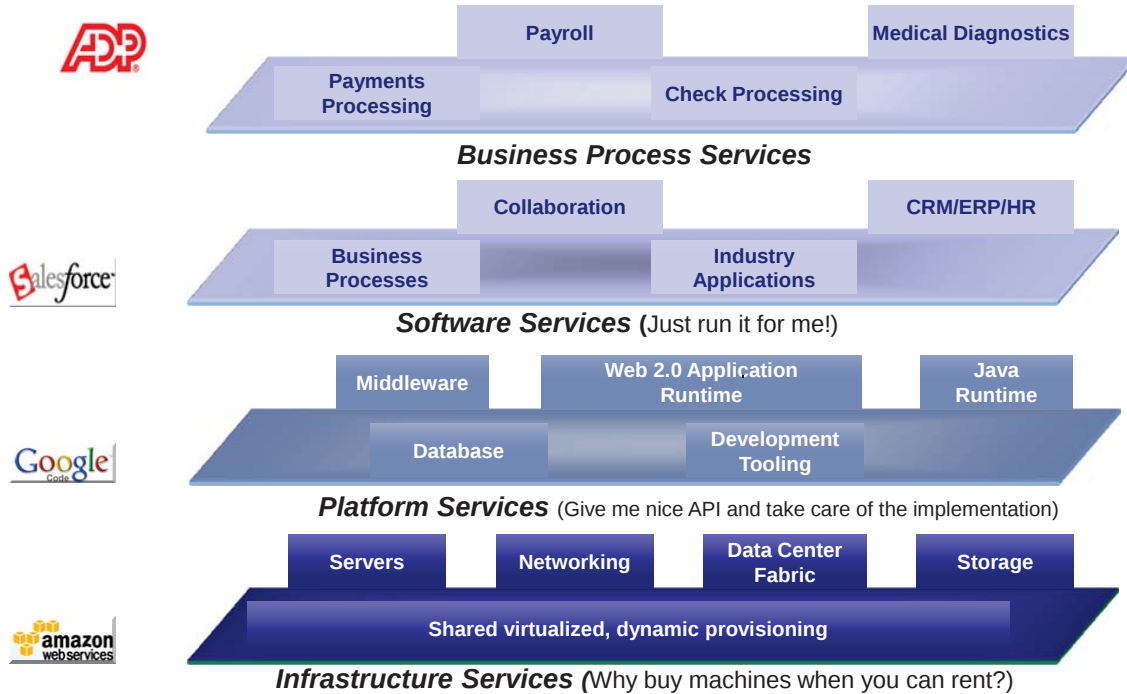
- Many many more ...



- What is Cloud Computing?
 - It is an emerging style of computing in which applications, data and IT resources are provided to users as services delivered over the network
 - It enables self-service, economies of scale and flexible sourcing options
 - “Cloud” refers to large Internet services like Google, Yahoo, IBM, etc that run on 10,000’s of machines
- Cloud Computing Essential Characteristics
 - ✓ On-demand self-service
 - ✓ Broad network access
 - ✓ Resource pooling
 - ✓ Rapid elasticity
 - ✓ Measured service

Cloud Computing Service Models

“Why do it yourself if you can pay someone to do it for you?”



➔ What Exactly Is Apache Hadoop ?

- A framework for running applications (aka jobs) on large clusters built on **commodity hardware** capable of processing **peta bytes of data**.
- A framework that transparently provides applications both reliability and data motion. It ensures **data locality**.
- It implements a computational paradigm named **Map/Reduce**, where the application is divided into self contained units of work, each of which may be executed or re-executed on any node in the cluster.
- It provides a distributed file system (**HDFS**) that stores data on the compute nodes, providing very high aggregate bandwidth across the cluster. HDFS is a massively distributed file system designed to run on cheap commodity hardware.
- Node **failures are automatically handled** by the framework.

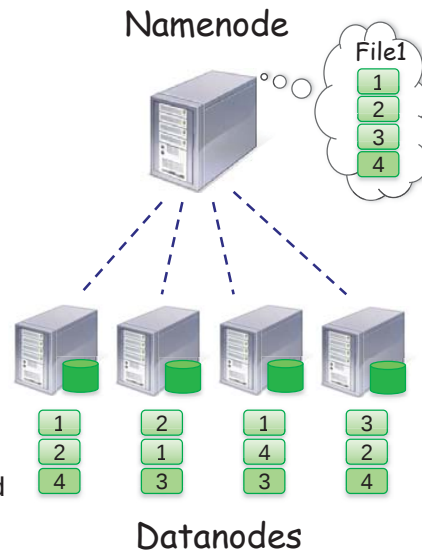
Hadoop Components

→ Distributed file system (HDFS)

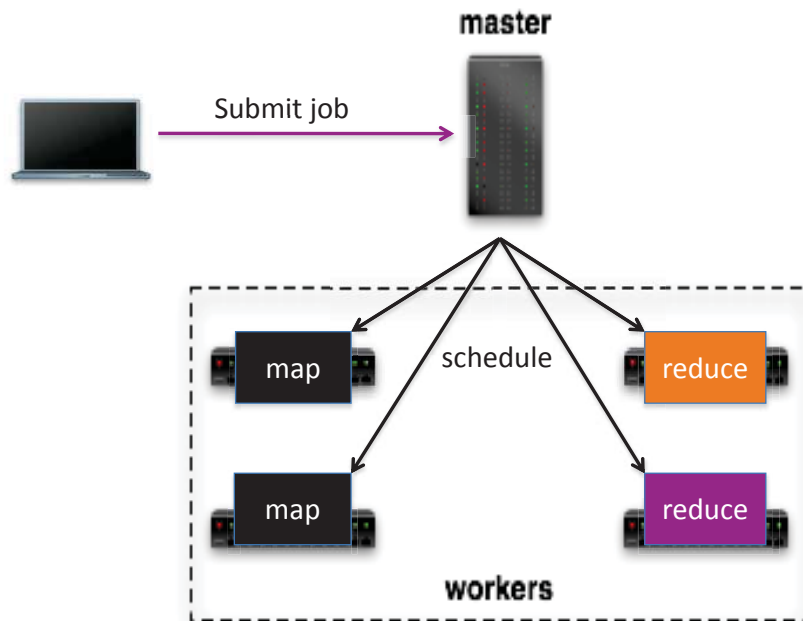
- Single namespace for entire cluster stores metadata (file names, block locations, etc)
- Each file is chopped up into a number of blocks (128MB)
- Fault tolerance is achieved by replicating these data blocks over a number of nodes (replicates data 3x for fault-tolerance)
- Optimized for large files, sequential reads
- Files are append-only

→ MapReduce framework

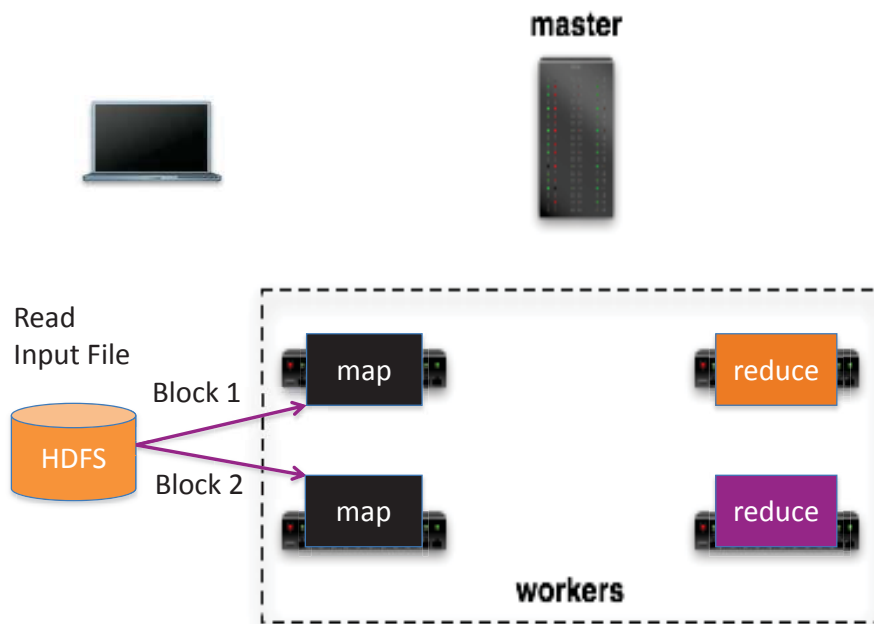
- Simple data-parallel programming model designed for scalability, work distribution and fault-tolerance
- Executes user jobs specified as “map” and “reduce” functions
- Pioneered by Google, processing 20 petabytes of data per day
- Popularized by open-source Hadoop project
- Used at Yahoo!, Facebook, Amazon, ...



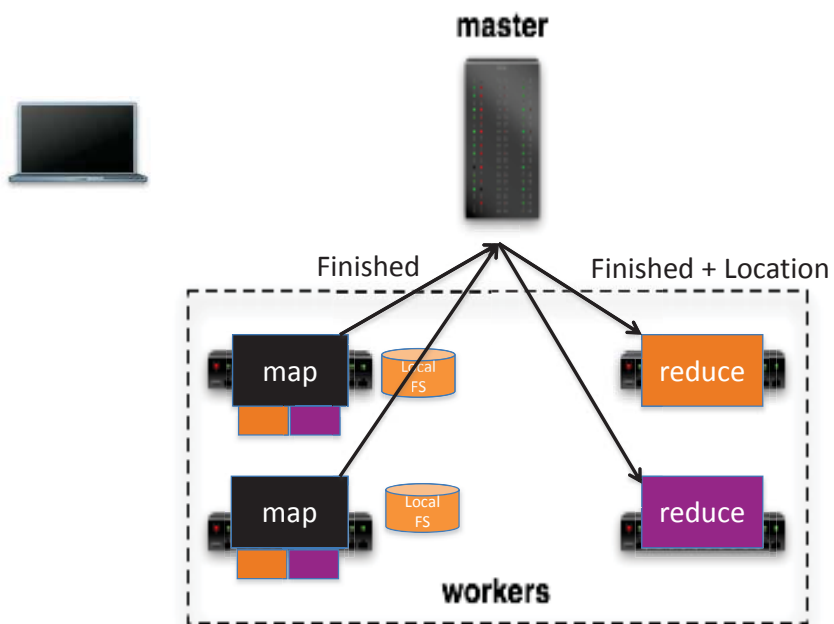
Dataflow in Hadoop



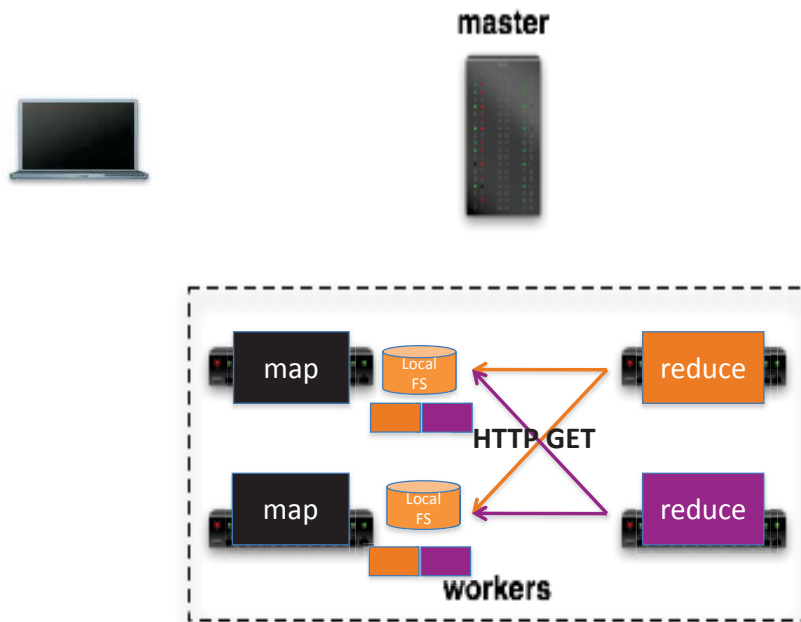
Dataflow in Hadoop



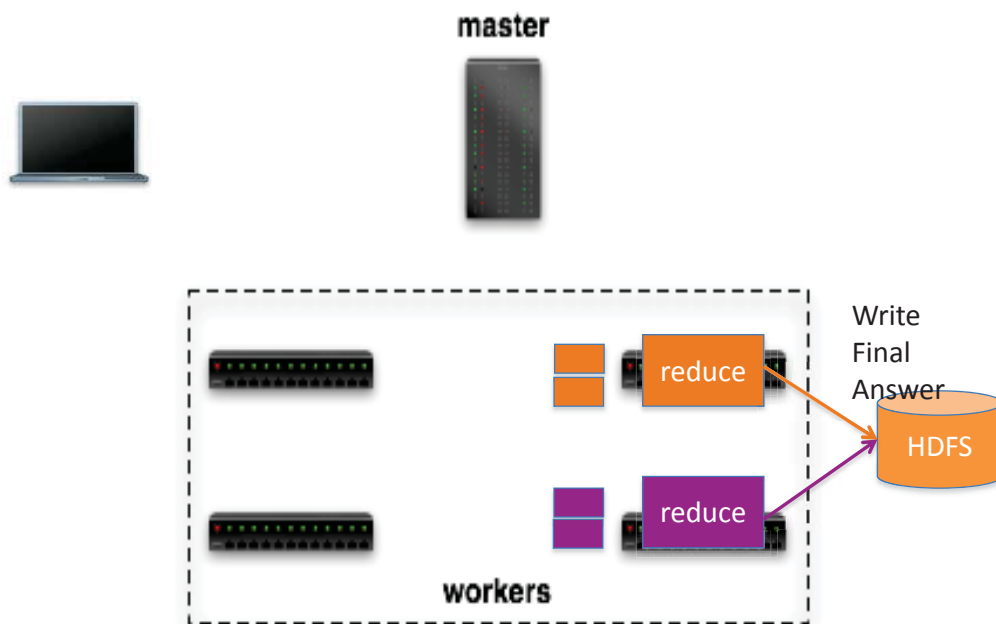
Dataflow in Hadoop



Dataflow in Hadoop



Dataflow in Hadoop



Takeaways

- ➔ By providing a data-parallel programming model, MapReduce can control job execution in useful ways:
 - Automatic division of job into tasks
 - Automatic placement of computation near data
 - Automatic load balancing
 - Recovery from failures & stragglers
- ➔ User focuses on application, not on complexities of distributed computing

Customer Entity Cleansing

Need for Data Quality



Critical Problems

- Need to create & maintain 360 degree views of customers, suppliers, products, locations, events
- Need to leverage data - make reliable decisions, comply with regulations, meet service agreements

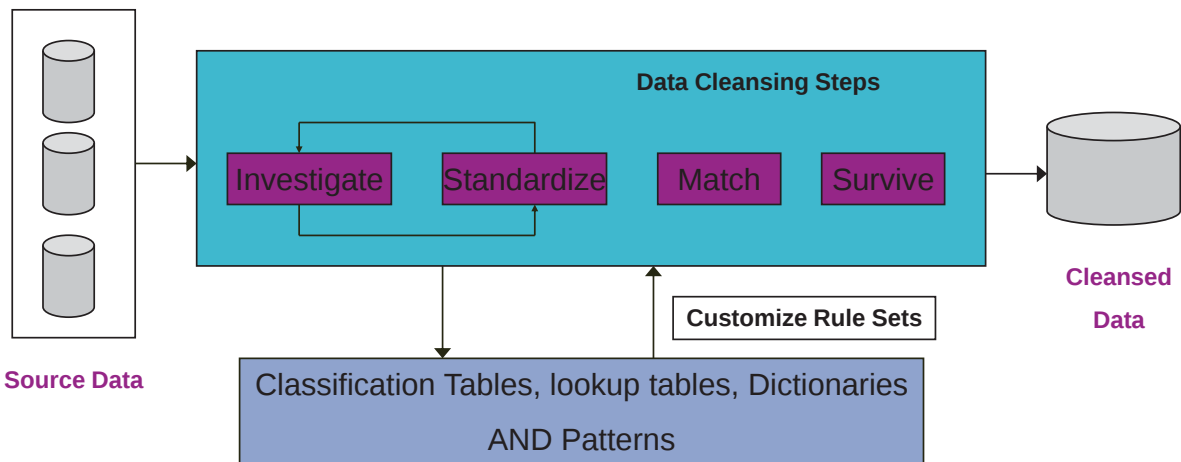
Why?

- No common standards across organization
- Unexpected values stored in fields
- Required information buried in free-form fields
- Fields evolve - used for multiple purposes
- No reliable keys for consolidated views
- Operational data degrades 2% per month

Approach

- Data Cleansing

Data Cleansing Steps



"9005 Essex Bldg" J Carpenter Fway Dallas Texas 75257"
 House Number Street City State Zip

Standardize "Fway" and "Bldg" to Freeway and Building and area names like
 "J Carpenter" to John Carpenter



Investigation – Word

123 St. Virginia St.

Parsing:

123 | St. | Virginia | St.

Separating multi-valued fields into individual pieces

Lexical analysis:

Number Street Alpha Street
Type Type
123 | St. | Virginia | St.

Determining business significance of individual pieces

Context Sensitive:

House Street Street
Number Name Type
123 | St. Virginia | St.

Identifying various data structures and content

“The instructions for handling the data are inherent within the data itself.”



Matching -- What Constitutes a Good Match?

Which of the following record pairs is a match? And how do you know?

W	HOLDEN	12	MAIN	ST
W	HOLDEN	12	MAINE	ST

W	HOLDEN	128	MAIN	PL	02111	12/8/62
W	HOLDEN	128	MAINE	PL	02110	12/8/62

WM	HOLDEN	128A	MAIN	SQ	02111	12/8/62	338-0824
WILL	HOLDEN	128A	MAINE	SQ	02110	12/8/62	338-0824

- ú Do you compare all the shared or common fields?
- ú Do you give partial credit?
- ú Are some fields (or some values) more important to you than others? Why?
- ú Do more fields increase your confidence?
- ú By how much? What is enough?



Two Methods to Decide a Match

Are these two records a match?

WILLIAM J	HOLDEN	128	MAIN	ST	02111	12/8/62			
WILLIAM JOHN	HOLDEN	128	MAINE	AVE	02110	12/8/62			
B	B	A	A	B	D	B	A	=	BBAABDBA
+5	+2	+20	+3	+4	-1	+7	+9	=	+49

Deterministic Decisions Tables:

- Fields are compared
- Letter grade assigned
- Combined letter grades are compared to a vendor delivered file
- Result: Match; Fail; Suspect

Probabilistic Record Linkage:

- Fields are evaluated for degree-of-match
- Weight assigned: represents the **"information content"** by value
- Weights are summed to derived a total score
- Result: Statistical probability of a match



Enhancing Data Quality Using Enterprise/Web Data



Product Standardization

→ Extract and standardize product attributes from product descriptions

Product Descriptions: CANNON CLS T220 SHT SET HAMPTON PLAID F
Fisher-Price Look & Learn Binocular Gift Set – Rainforest Life FP34

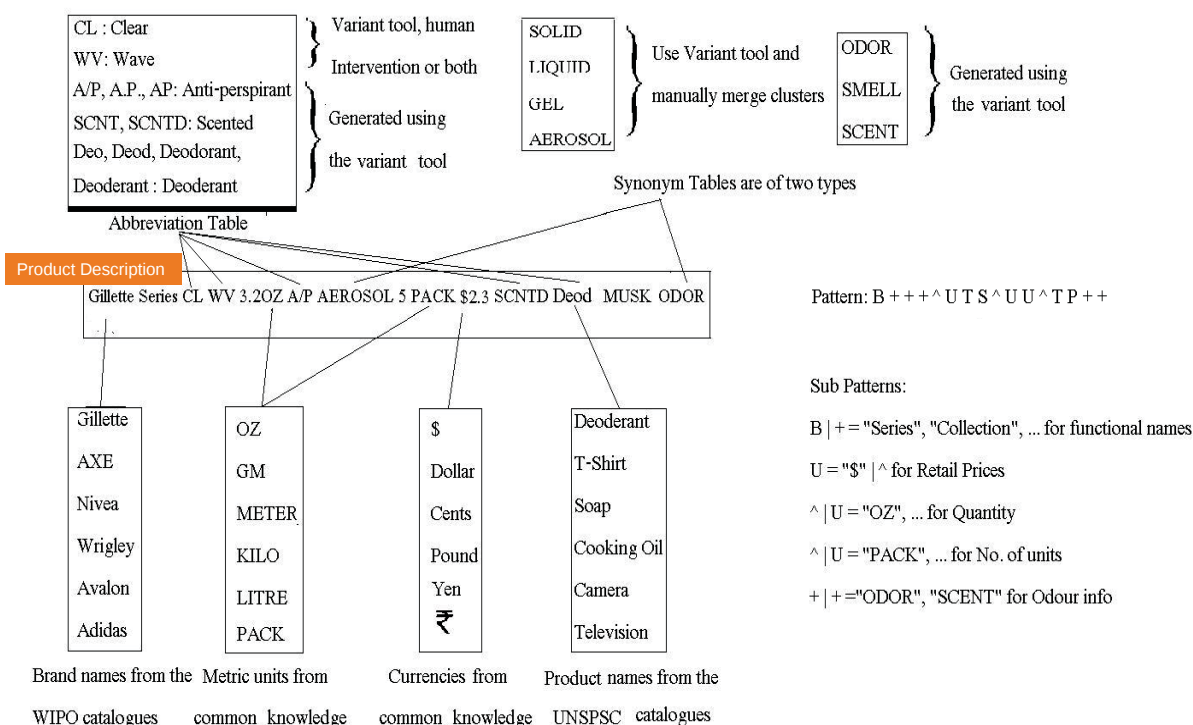
Extraction and Standardization:

Brand	Item No	Item Name	Model	Colour	Series
CANNON	CLS T220	SHIRT SET	-	HAMPTON PLAID	F
Fisher-Price	FP34	Look & Learn Binocular gift set	Rainforest Life	-	--

Product attributes are not standard

Classification and standardization based on the defined attributes

Attribute Extraction and Attribute Standardization





Thank you!



International Association for Information and Data Quality

Advancing the information
and data quality profession
and its body of knowledge

IAIDQ

- Founded in 2004
 - by Larry English and Tom Redman
- 350 financial members
- 33 countries in 6 continents
- Top 5:
 - USA
 - Canada
 - Australia
 - UK
 - China

Key products

- Professional certification
- Conference (USA)
- Conference support: EU, Americas, Asia, AU
- Quarterly newsletter
- Industry reports
- Webinars
- Monthly update



Find us at

iaidq.org