DATA7001
INTRODUCTION TO DATA SCIENCE

Module 4 Making the Data Confess

Module Topics

- Hindsight (search and query),
  - What happened?
- Insight (knowledge discovery)
  - Why is it happening?
- Foresight (prediction)
  - What will happen?

→ What should happen (making it 'actionable')

Task and Discussion

Give an example for each type of data analytics

Global Health Example

Hindsight:

Insight:

Foresight:

Hindsight

- Relational queries on transactional databases
- Aggregation queries on data warehouses
- Informational retrieval and text analytics

Relational query (more in INFS7901)

SUPPLIER [Sno, Sname, Addr]

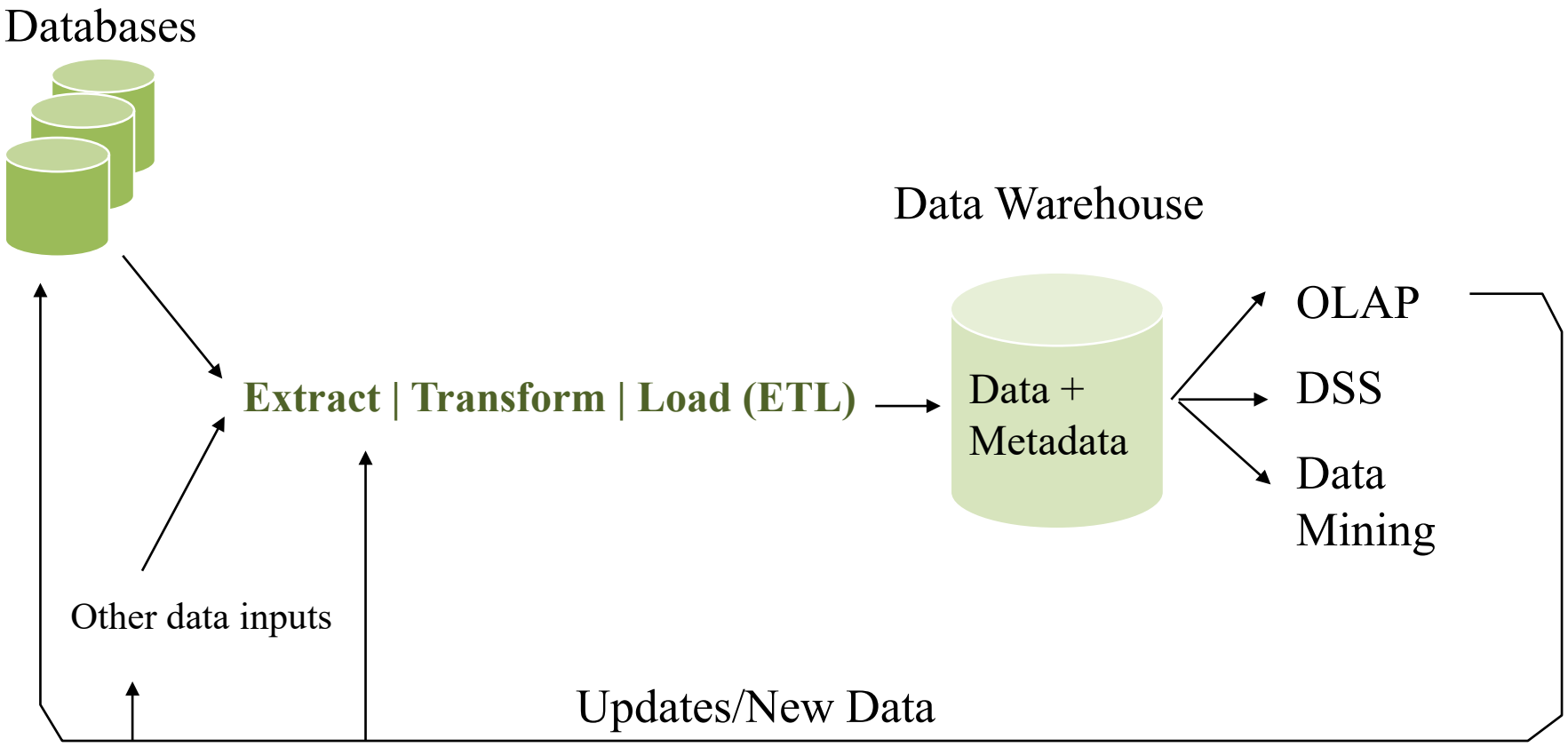
PART [Pno, Pname, Price]

SHIPMENT [Sno, Pno, Qty]

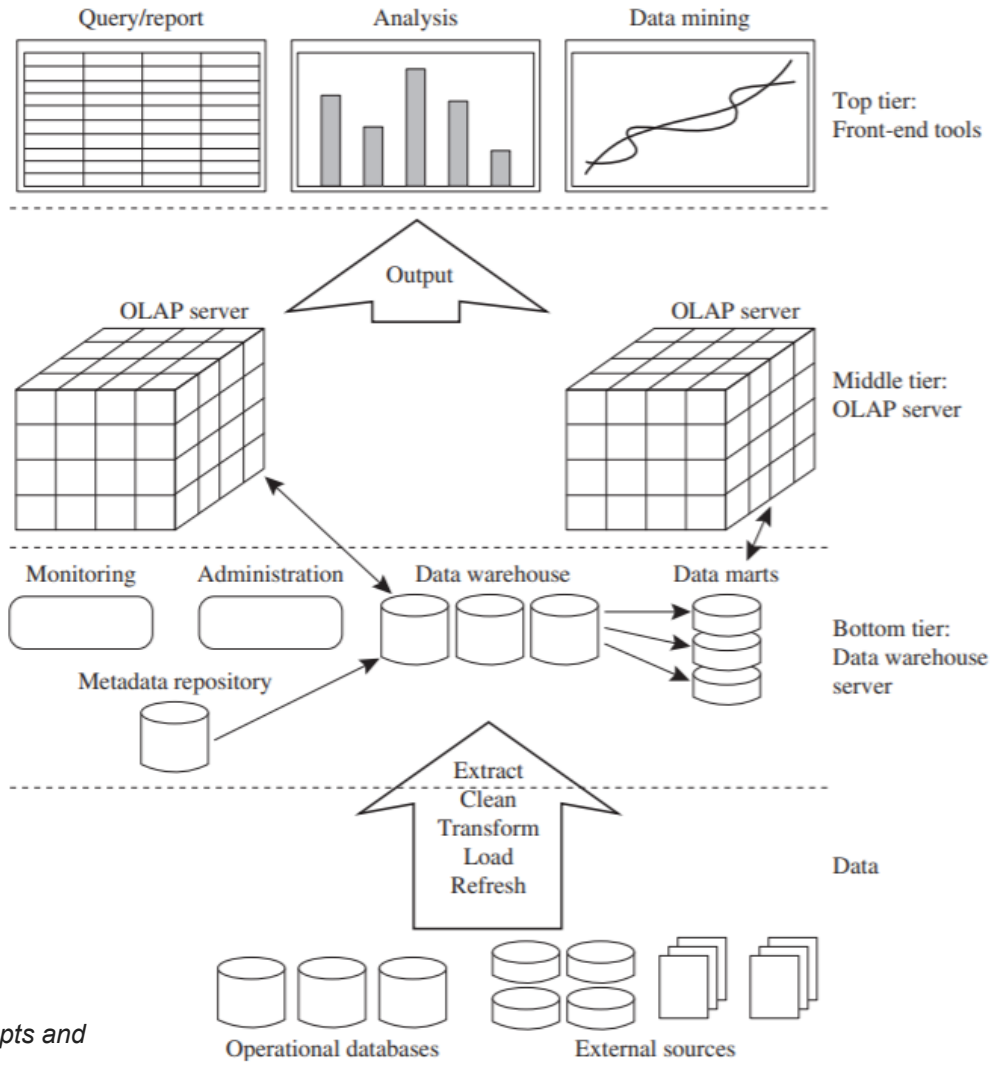
Find all parts of price more than $1000, supplied by supplier 'Big Company'

```
SELECT Pno
FROM SHIPMENT
WHERE Pno IN
    (SELECT Pno)
    FROM PART
    WHERE Price > $1000)
AND Sno IN
    (SELECT Sno
    FROM SUPPLIER
    WHERE Sname = 'Big
                Company')
```

Data warehousing (more in INFS7907)



Three-layer Data Warehousing Architecture



Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

Why Data Warehouse?

- Applications consist both of updates and queries. Some queries are large scale aggregation reports.
- Both updates and queries must lock data resources. Large scale aggregation reports lock many resources for a long time.
- If high frequency of updates coincides with high frequency of reports, there is competition for computing resources.

For example, student enrolment transactions at beginning of semester coincide with high report demand for determining whether room sizes, tutor allocations, etc are adequate.

Other examples:
Sales and inventory (supermarket)
Public transport (Translink)
Social services (Centrelink)
…

Multi-dimensional model

Distinct from relational model

R [A1, A2 … An]

Consists of a *Fact Table* and multiple *Dimension Tables.*

- Fact table: numeric measures such as **amounts sold** and **#items sold** with their dimension id.
- Dimension Tables: categorical measures such as **branch**, **location**, **time** along with their ids (each dimension is one table).

Example: Fact table for Sales Data

Key			Facts
Day	Product	Store	Sales (AUD)
9.2.04	Milk	Toowong	3412
10.2.04	Milk	Toowong	2918
9.2.04	Bread	Toowong	2918
10.2.04	Bread	Toowong	3445
9.2.04	Milk	Sunnybank	5440
10.2.04	Milk	Sunnybank	4992
9.2.04	Bread	Sunnybank	2918
10.2.04	Bread	Sunnybank	3067

Dimensions

- Each key is a dimension - example has three
- Dimensions have hierarchical organisation
  - Days grouped into weeks, months, quarters, years
  - Product groups aggregated
    - Milk -> dairy -> perishable -> food
    - Bread -> baked goods -> perishable -> food
  - Stores grouped into regions
    - Toowong -> Metro Brisbane -> QLD -> Australia -> Oceania
- Dimensions are organised by dimension tables

Dimension tables define the facts

- Each dimension is a projection of the fact table onto one of its keys.

Day
9.2.04
10.2.04

Product
Milk
Bread

Store
Toowong
Sunnybank

More general classes stored in dimension tables

Day	Month	Qtr	Year
9.2.04	Feb	1	2004
10.2.04	Feb	1	2004

Store	District	Region
Toowong	North	Brisbane
Sunnybank	South	Brisbane

Product	Kind	Type	Class
Milk	Dairy	Perishable	Food
Bread	Bakery	Perishable	Food

A *Star Schema*

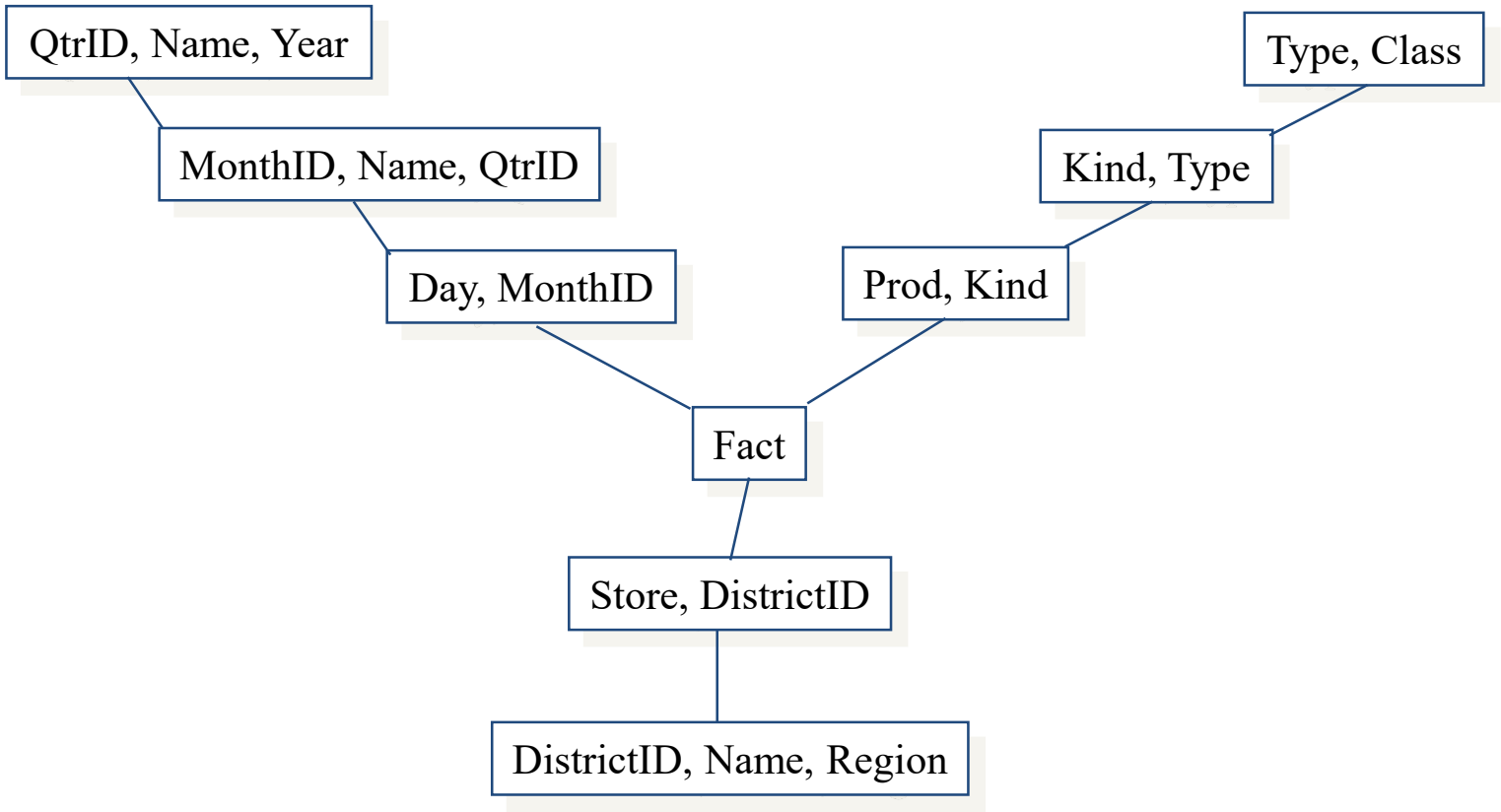
Day	Month	Qtr	Year
9.2.04	Feb	1	2004
10.2.04	Feb	1	2004

Store	District	Region
Toowong	North	Brisbane
Sunnybank	South	Brisbane

Facts

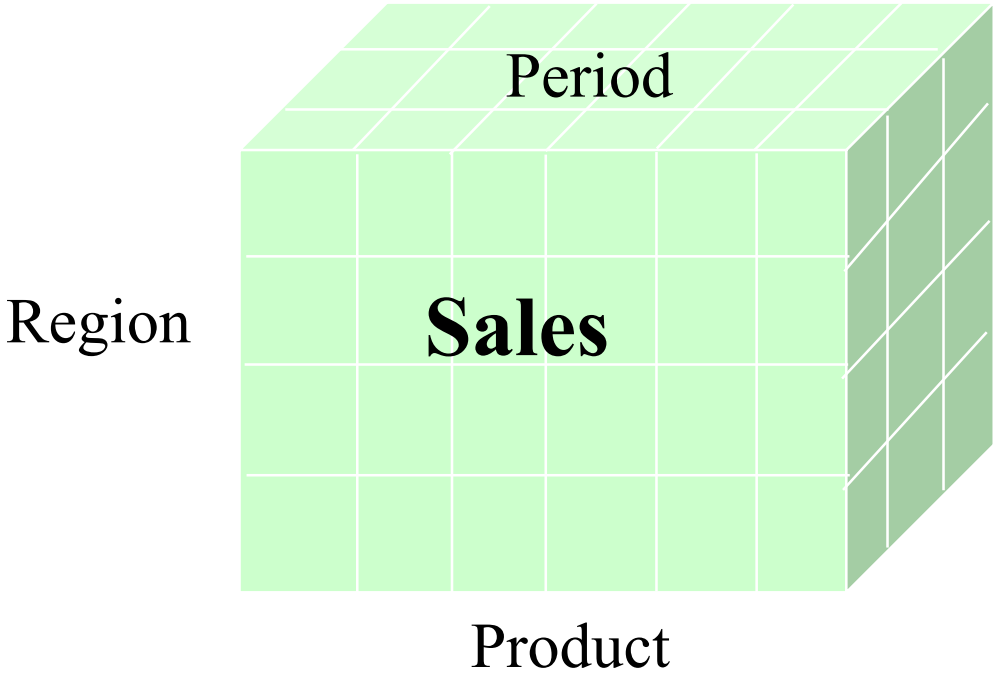
Product	Kind	Type	Class
Milk	Dairy	Perishable	Food
Bread	Bakery	Perishable	Food

Normalized dimension tables make a *Snowflake Schema*



Data Cube

- A popular three-dimensional model for a data warehouse
  - More than three dimensions result in Hypercubes



Online Analytical Processing (OLAP)

- Influenced by SQL and spreadsheets.

- A common operation is to aggregate a measure over one or more dimensions.
  - Find total sales.
  - Find total sales for each city, or for each state.
  - Find top five products ranked by total sales.

Typical OLAP Queries

- Roll-up
  - Aggregating at different levels of a dimension hierarchy.
- Drill-down
  - Disaggregate to a finer-grained view, inverse of roll-up
- Slice and dice
  - Perform select operations on the dimensions, similar to HAVING clause in SQL
- Pivoting (cross-tabulation)
  - Rotate data cube to show a different orientation of axes

Example roll-up

	Product		Total
Day	Milk	Bread	Perishables
9.2.04	8952	5836	14788
10.2.04	7910	8059	15969
	Product Group		Total
Day	Perishables	Canned Goods	All Groups
9.2.04	14788	55621	206771
10.2.04	15969	68123	310885

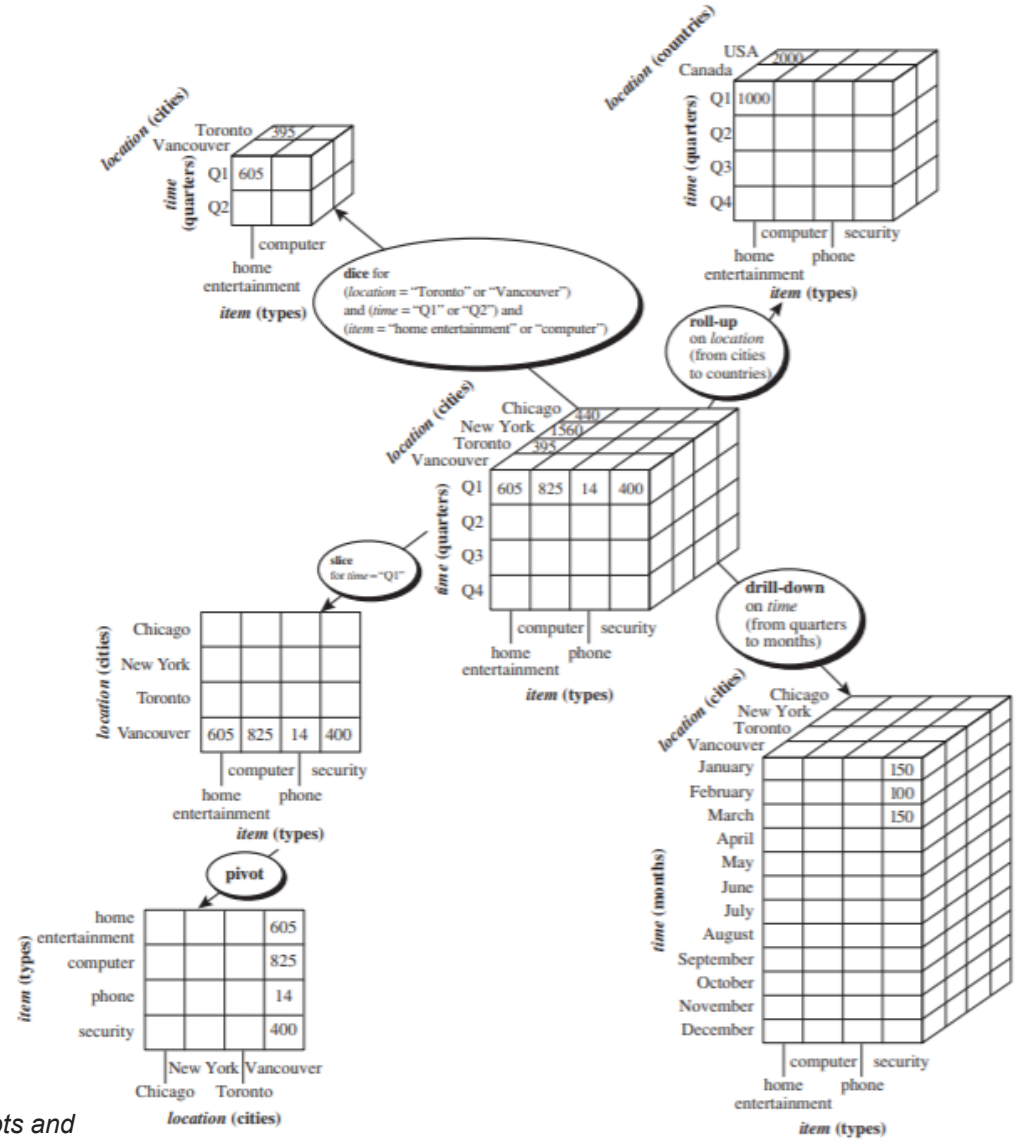
Roll up milk, bread to compare perishables with other product groups

Example drill down

	Product Group		Total
Day	Perishables	Canned Goods	All Groups
9.2.04	14788	55621	206771
10.2.04	15969	68123	310885
	Product		Total
Day	Milk	Bread	Perishables
9.2.04	8952	5836	14788
10.2.04	7910	8059	15969

Drill down perishables to constituent products

OLAP operations



Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

OLAP and Decision Support

- OLAP queries are typically aggregate queries.
  - Expressing in SQL is hard, and hence goal is to give non-SQL users the tools for a select number of popular aggregate queries
  - Pre-computation is essential for interactive response times
- Data warehouses are a collection of asynchronously created views of transactional databases for read-only aggregate queries
  - Frequency of refreshing the data warehouse has to be determined in accordance with requirements decision support

Task and Discussion

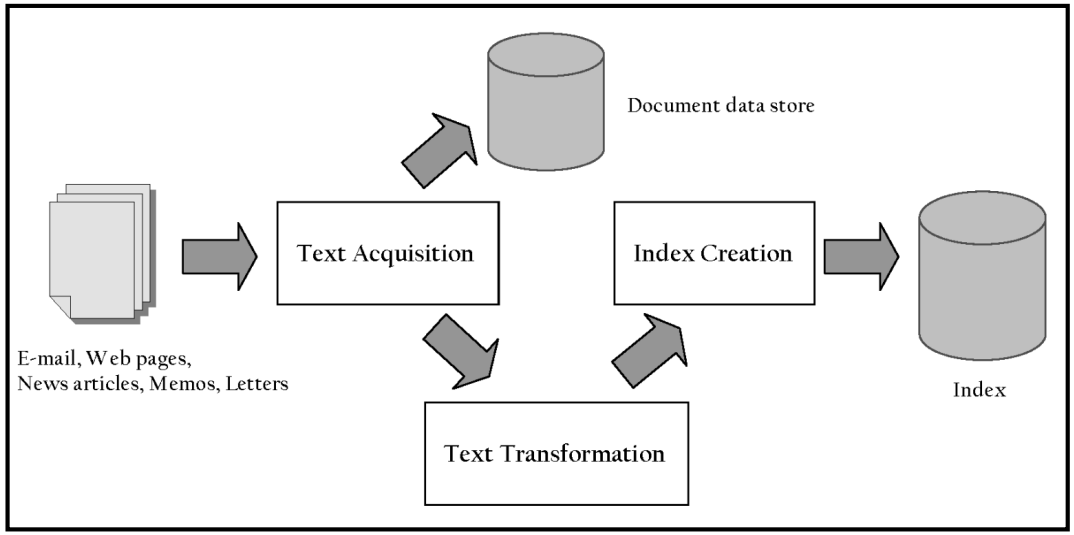
Give example of a Data Warehouse outlining the fact and dimension tables and 3 OLAP queries

Information Retrieval (more in INFS7410)

- Text Search
  - Search engines
- Text Analytics
  - Documents (text corpus)
  - Short text (social media)

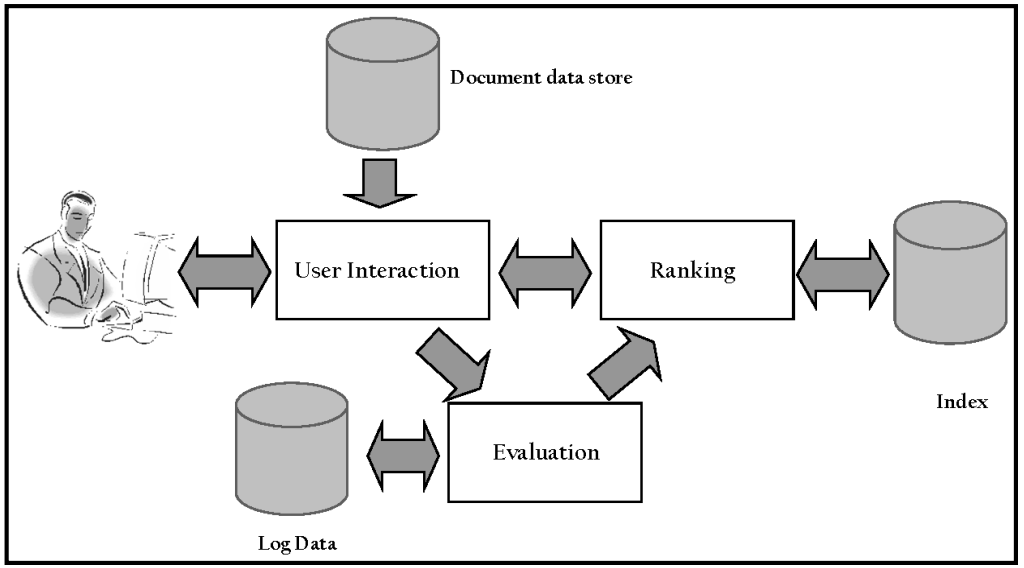
- Natural Language Processing (NLP)

Search Engine Architecture



Indexing Process

Query Process



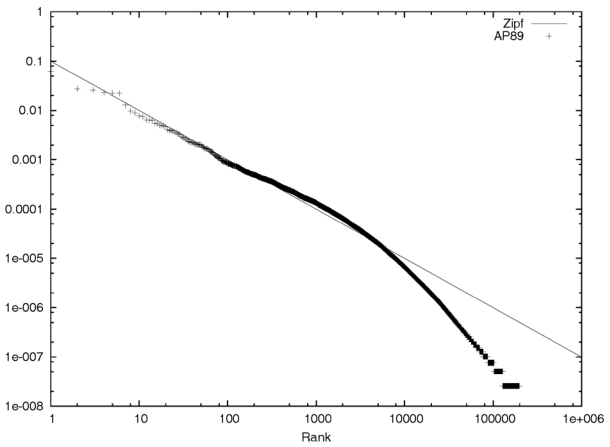
Text Search

- General Web Search Engine
  - Google, Bing, Yahoo, Yandex, Baidu
- Vertical Search Engine
  - Enterprise search (e.g. search in a website content)
  - Special purpose search (e.g. Google Scholar, PubMed).

Text Statistics

- Distribution of word frequencies is very skewed
- A few words occur very often, many words hardly ever occur
  - e.g., two most common words (“the”, “of”) make up about 10% of all word occurrences in text documents
- Retrieval models and ranking algorithms depend heavily on statistical properties of words
  - e.g., important words occur <u>often in documents</u> but are <u>not high frequency in collection</u> (Zipf’s Law)

Word	*Freq.*	*r*	P_r(%)	$r.P_r$	*Word*	*Freq*	*r*	P_r(%)	$r.P_r$
the	2,420,778	1	6.49	0.065	has	136,007	26	0.37	0.095
of	1,045,733	2	2.80	0.056	are	130,322	27	0.35	0.094
to	968,882	3	2.60	0.078	not	127,493	28	0.34	0.096
a	892,429	4	2.39	0.096	who	116,364	29	0.31	0.090
and	865,644	5	2.32	0.120	they	111,024	30	0.30	0.089
in	847,825	6	2.27	0.140	its	111,021	31	0.30	0.092
said	504,593	7	1.35	0.095	had	103,943	32	0.28	0.089
for	363,865	8	0.98	0.078	will	102,949	33	0.28	0.091
that	347,072	9	0.93	0.084	would	99,503	34	0.27	0.091
was	293,027	10	0.79	0.079	about	92,983	35	0.25	0.087
on	291,947	11	0.78	0.086	i	92,005	36	0.25	0.089
he	250,919	12	0.67	0.081	been	88,786	37	0.24	0.088
is	245,843	13	0.65	0.086	this	87,286	38	0.23	0.089
with	223,846	14	0.60	0.084	their	84,638	39	0.23	0.089
at	210,064	15	0.56	0.085	new	83,449	40	0.22	0.090
by	209,586	16	0.56	0.090	or	81,796	41	0.22	0.090
it	195,621	17	0.52	0.089	which	80,385	42	0.22	0.091
from	189,451	18	0.51	0.091	we	80,245	43	0.22	0.093
as	181,714	19	0.49	0.093	more	76,388	44	0.21	0.090
be	157,300	20	0.42	0.084	after	75,165	45	0.20	0.091
were	153,913	21	0.41	0.087	us	72,045	46	0.19	0.089
an	152,576	22	0.41	0.090	percent	71,956	47	0.19	0.091
have	149,749	23	0.40	0.092	up	71,082	48	0.19	0.092
his	142,285	24	0.38	0.092	one	70,266	49	0.19	0.092
but	140,880	25	0.38	0.094	people	68,988	50	0.19	0.093



Top 50 Words and Zipf’ s Law for AP89 (collection of Associated Press articles from 1989)

Text Search

- Many technical challenges
  - Estimating result size
  - Text analysis (Tokenization, Stopping, Stemming, Similarity, …) and Indexing
  - Measures for ranking effectiveness (precision, recall, f-measure, …) and efficiency (processor time, latency, throughput, index size, …)
  - Evaluation (objective vs subjective measures, and significance)

Text Analytics

Extracting meaning from text corpus:

- **Text classification**: sentiment analysis
- **Text clustering**: News clustering in GNews
- **Named Entity Recognition**: detection of names of people, locations, and organisations.
- **Machine Translation**

Text Analytics and Ambiguity

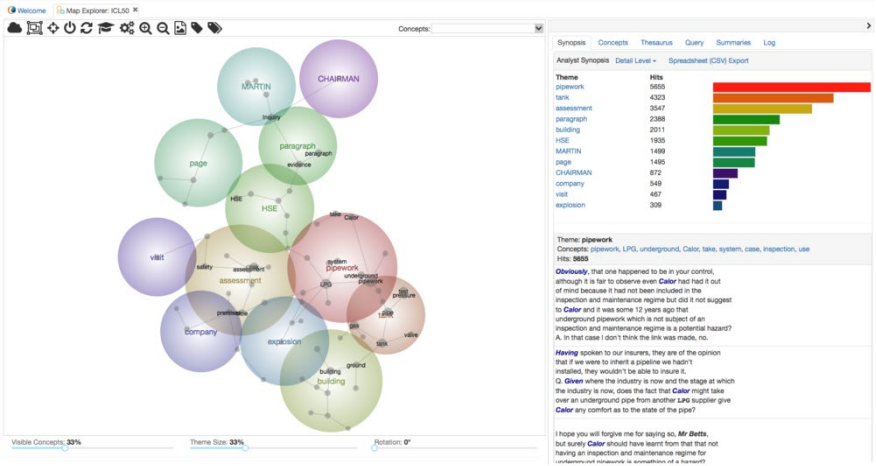
Extracting meaning from text corpus

What is *Tiger*?

Hard problems: Entity extraction, disambiguation, and linking
A little easier with the use of context: “The **tiger** was delicious.”

Tool for Text Analytics

- https://www.leximancer.com
  - Started at UQ, Andrew Smith ISSR



- Solr, ElasticSearch (built on lucene)