Big data workloads and real-world data sets

Gang Lu

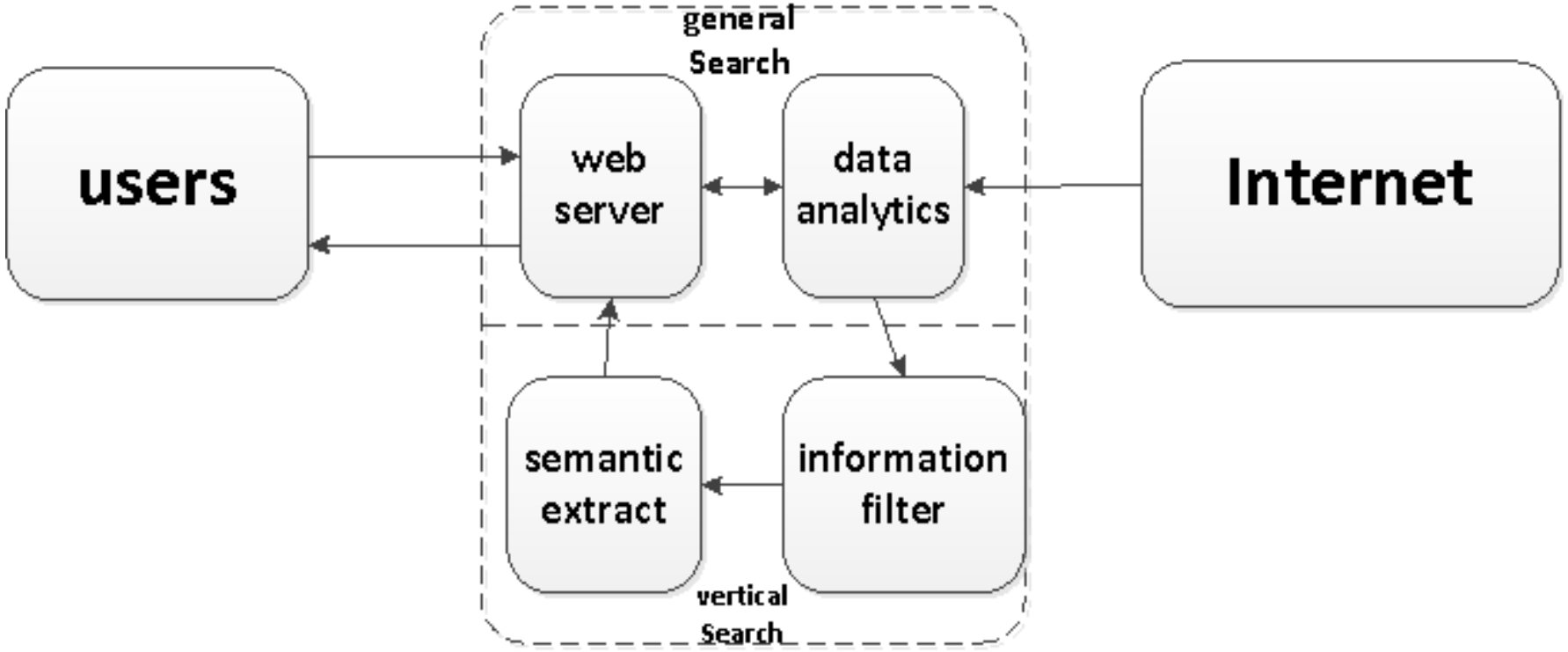
Beijing Academy of Frontier Science and Technology

BigDataBench Tutorial, ASPLOS 2016
Atlanta, GA, USA

Five domains

- Search engine
- Social network
- E-commerce
- Multi-media
- Bioinformatics

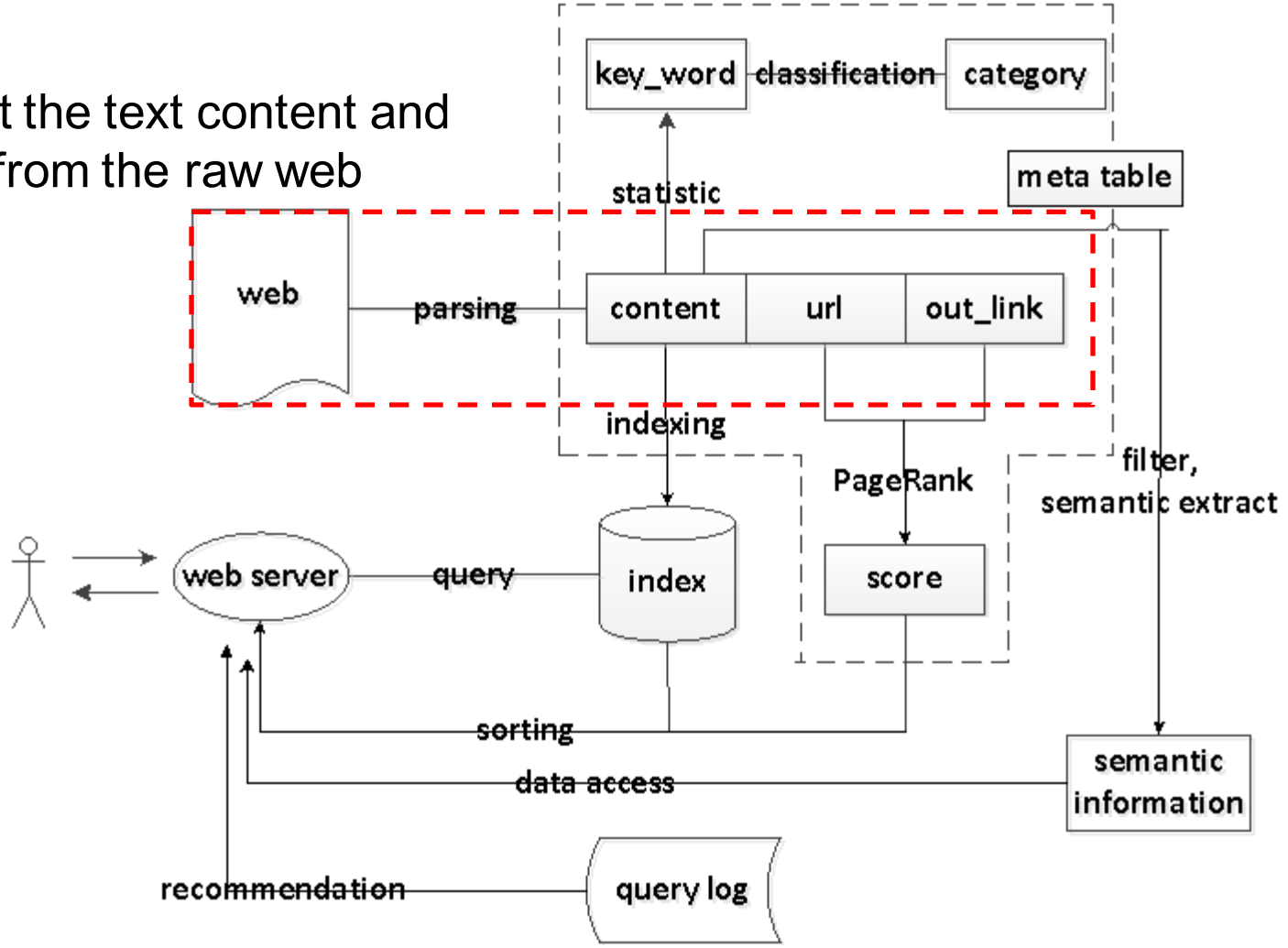
Search Engine



General search and vertical search
Online server and Offline analytics

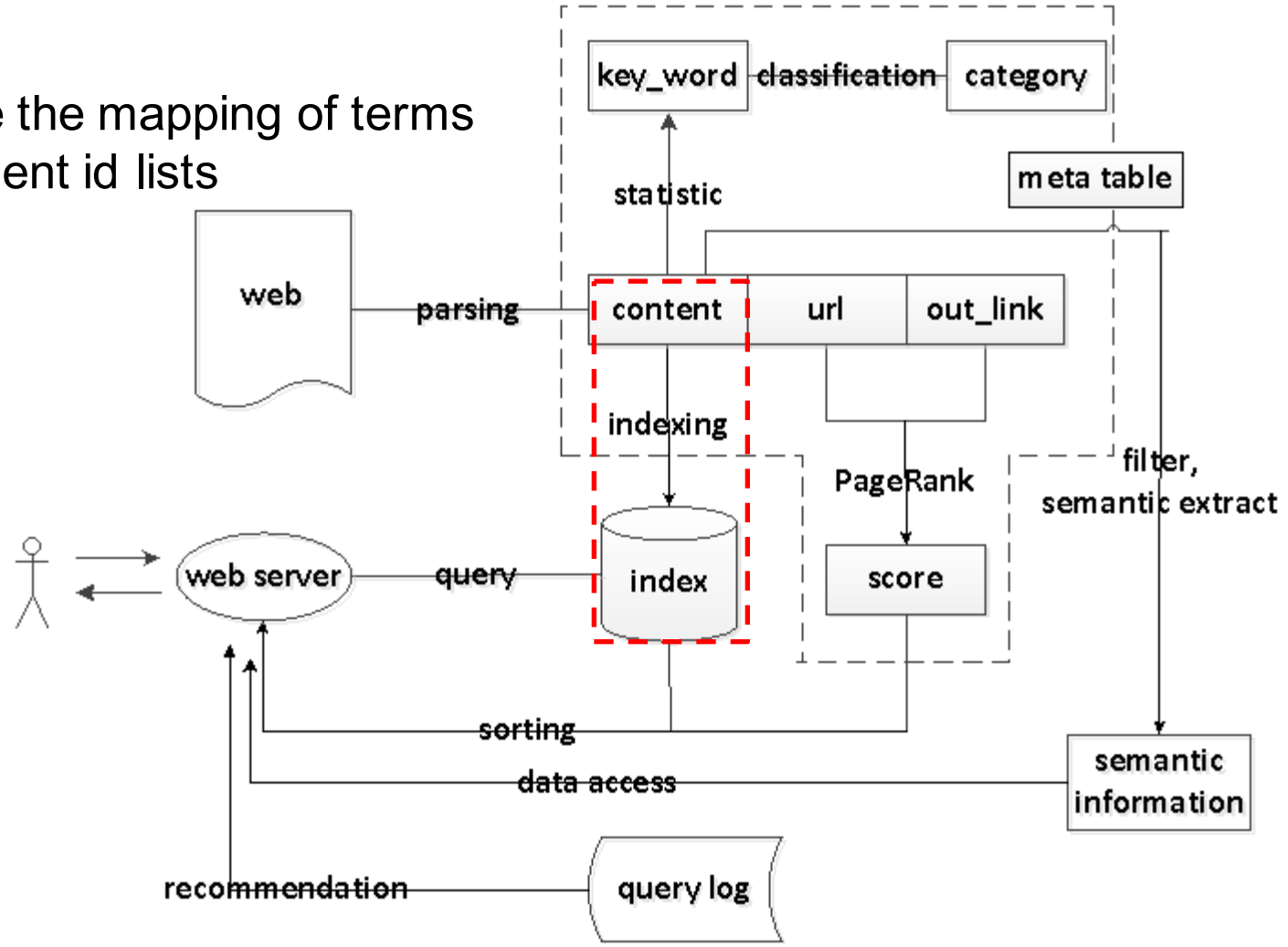
Search Engine: Parsing

- Parsing:
  - To extract the text content and out links from the raw web pages



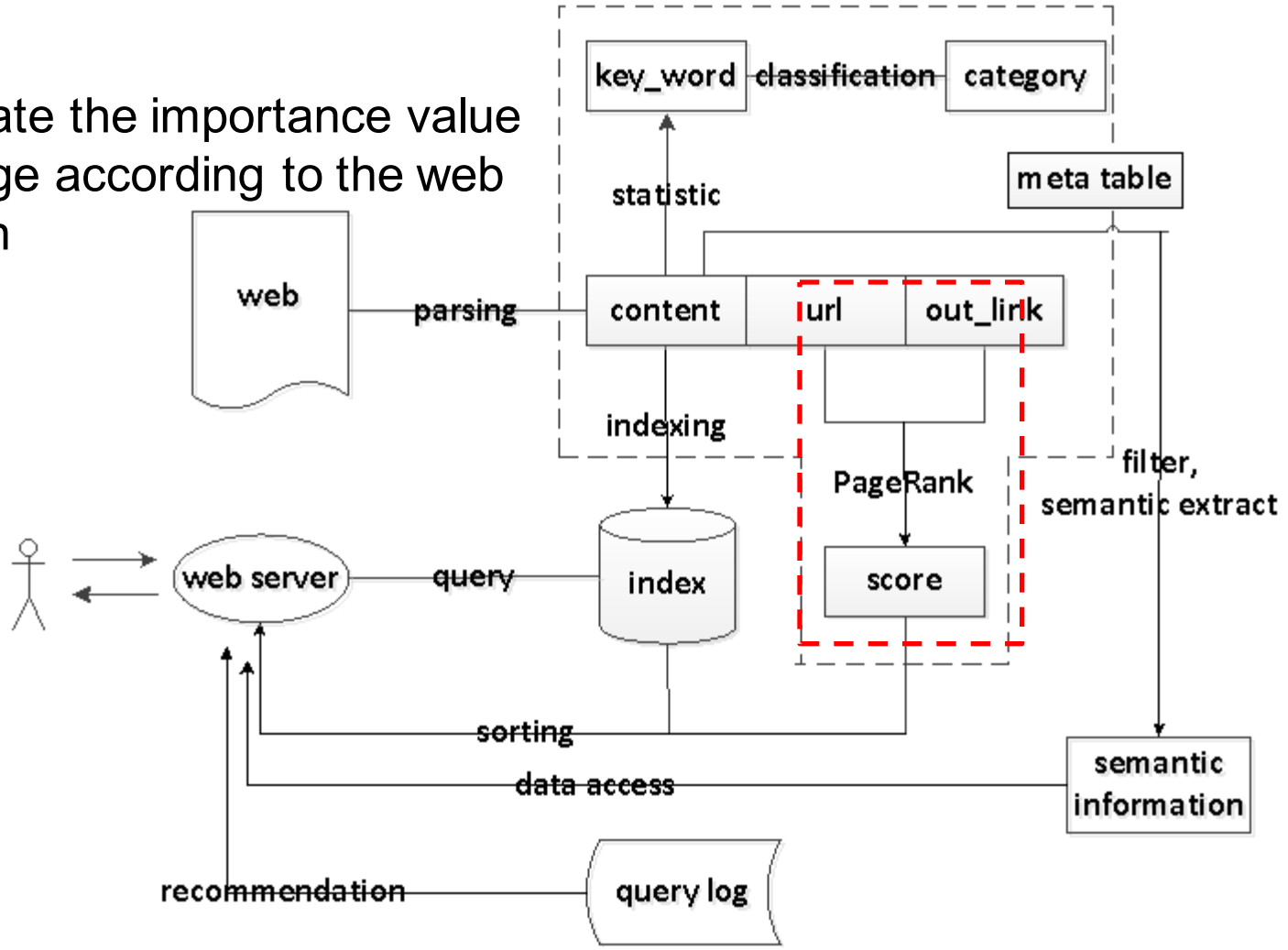
Search Engine: Indexing

- Indexing
  - To create the mapping of terms to document id lists



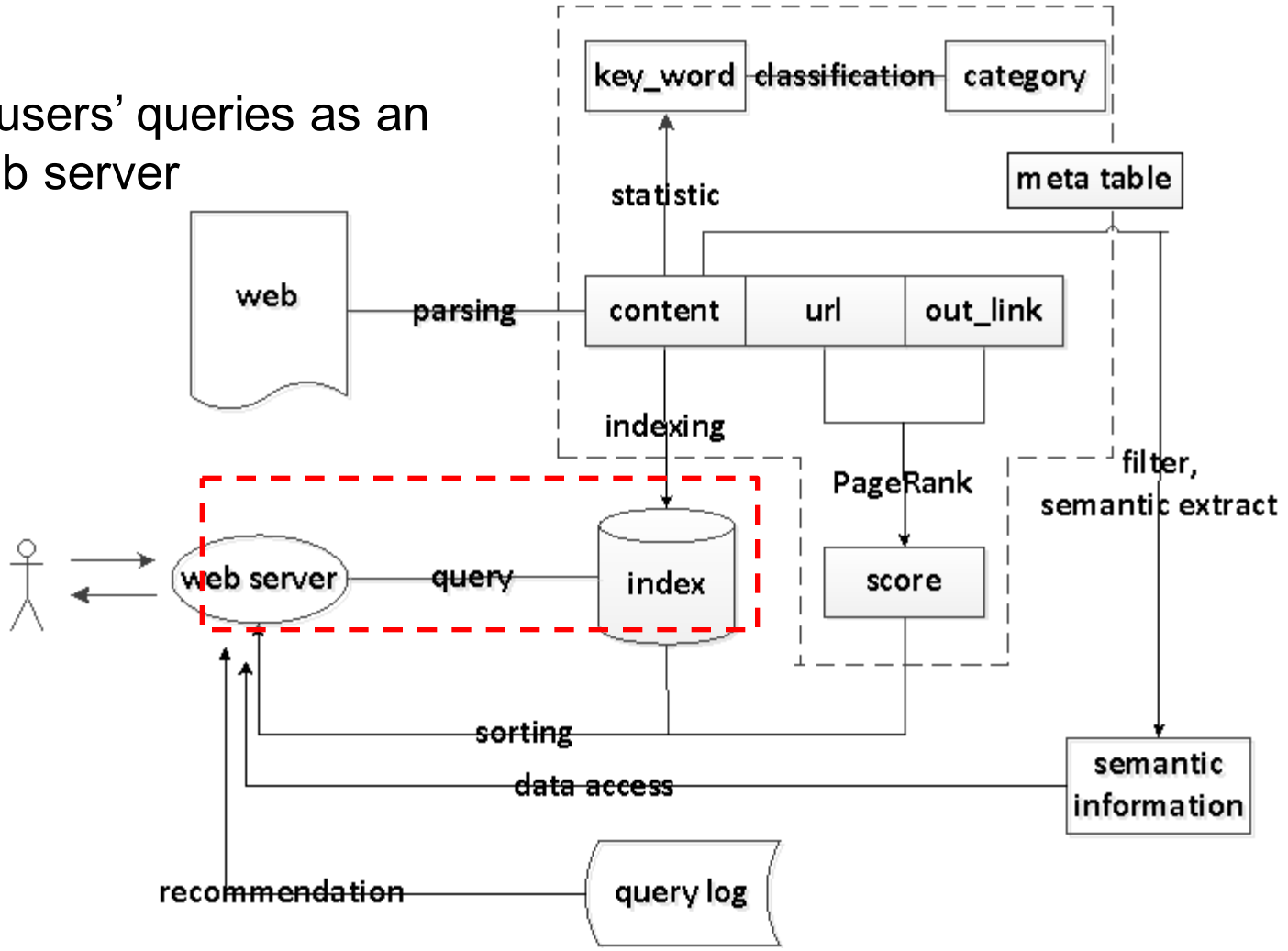
Search Engine: PageRank

- PageRank
  - To calculate the importance value of the page according to the web link graph



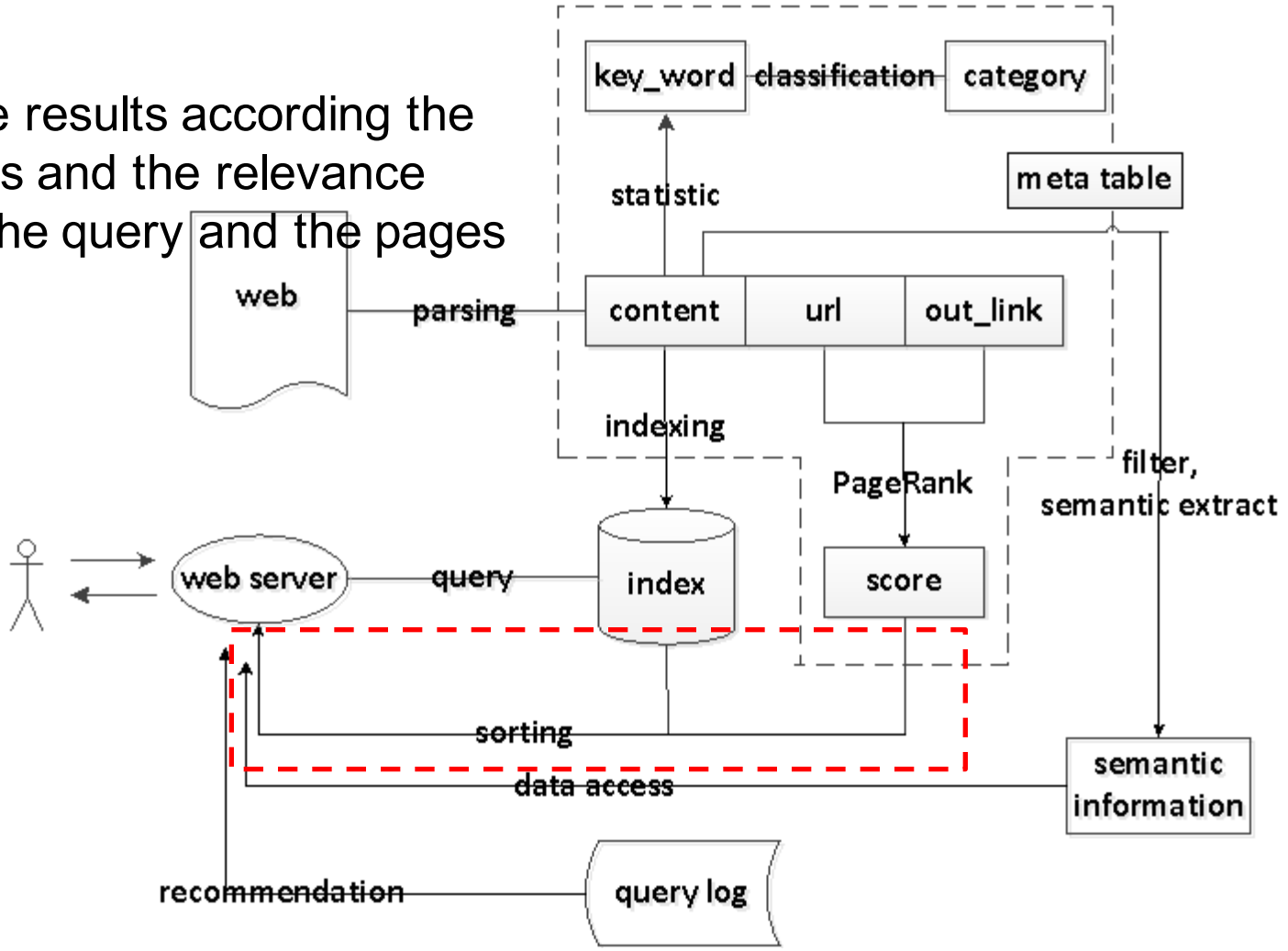
Search Engine: Search query

- Querying
  - To serve users' queries as an online web server



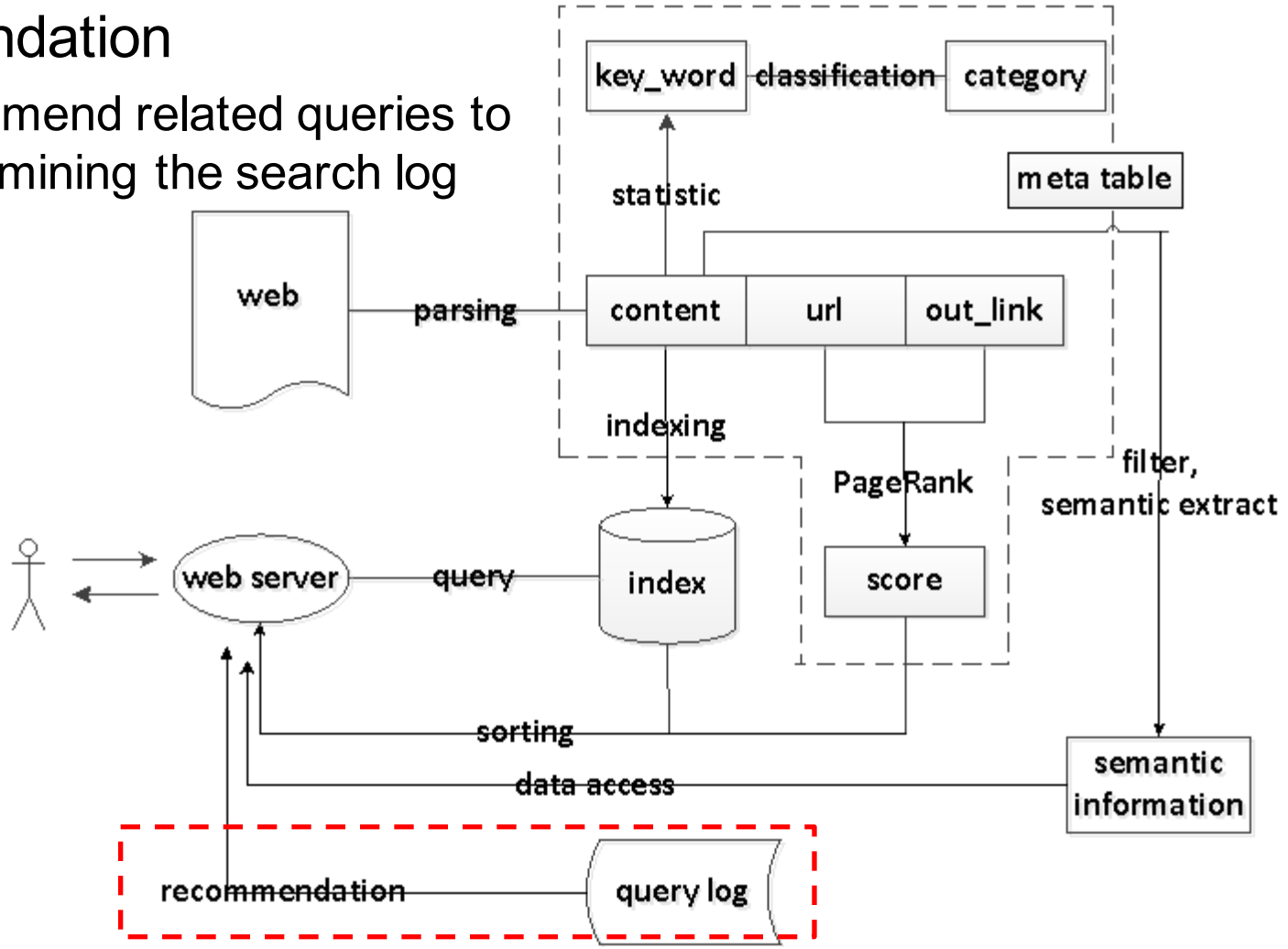
Search Engine: Sorting

- Sorting
  - To sort the results according the page ranks and the relevance between the query and the pages



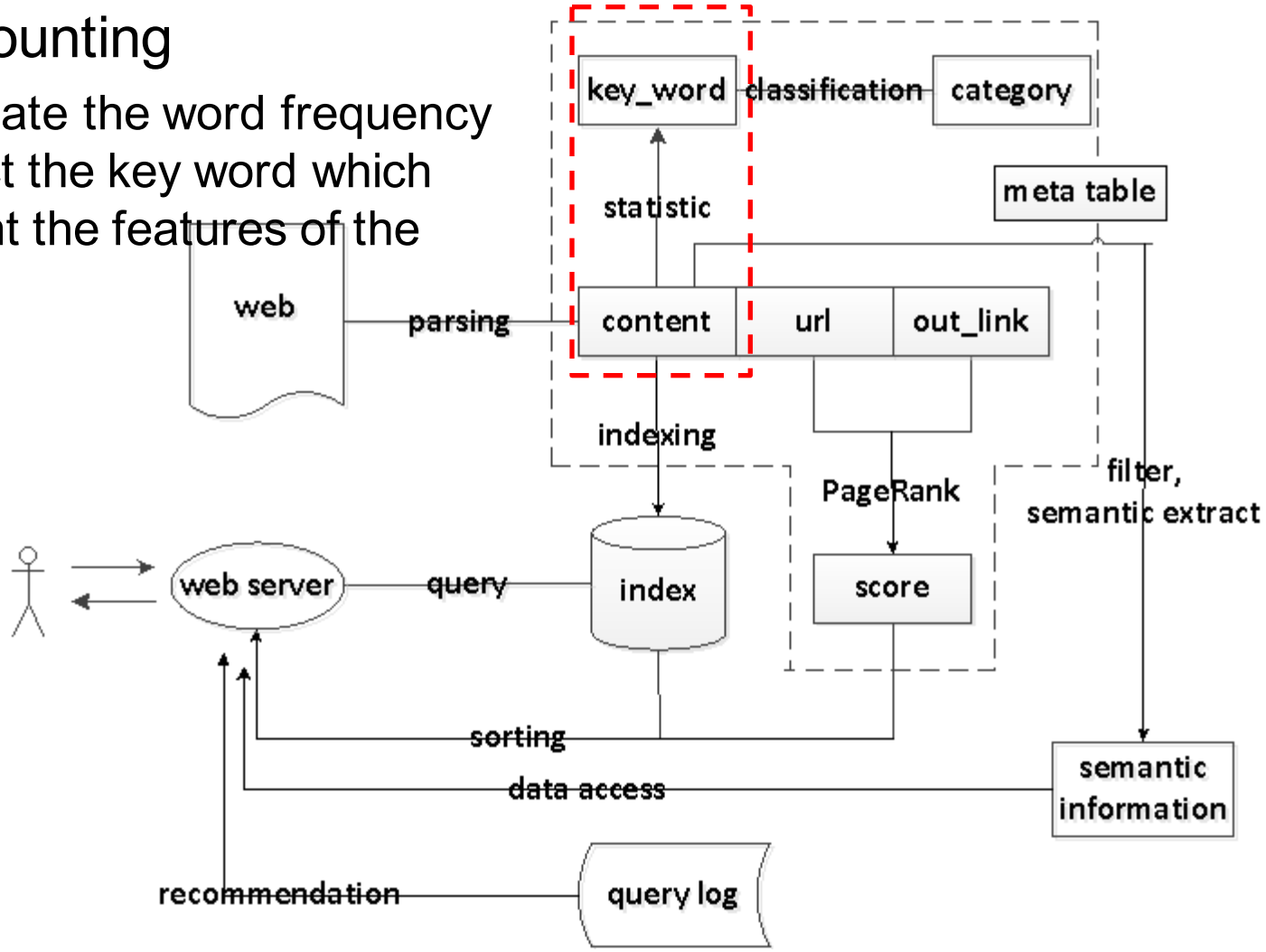
Search Engine: Recommendation

- Recommendation
  - To recommend related queries to users by mining the search log



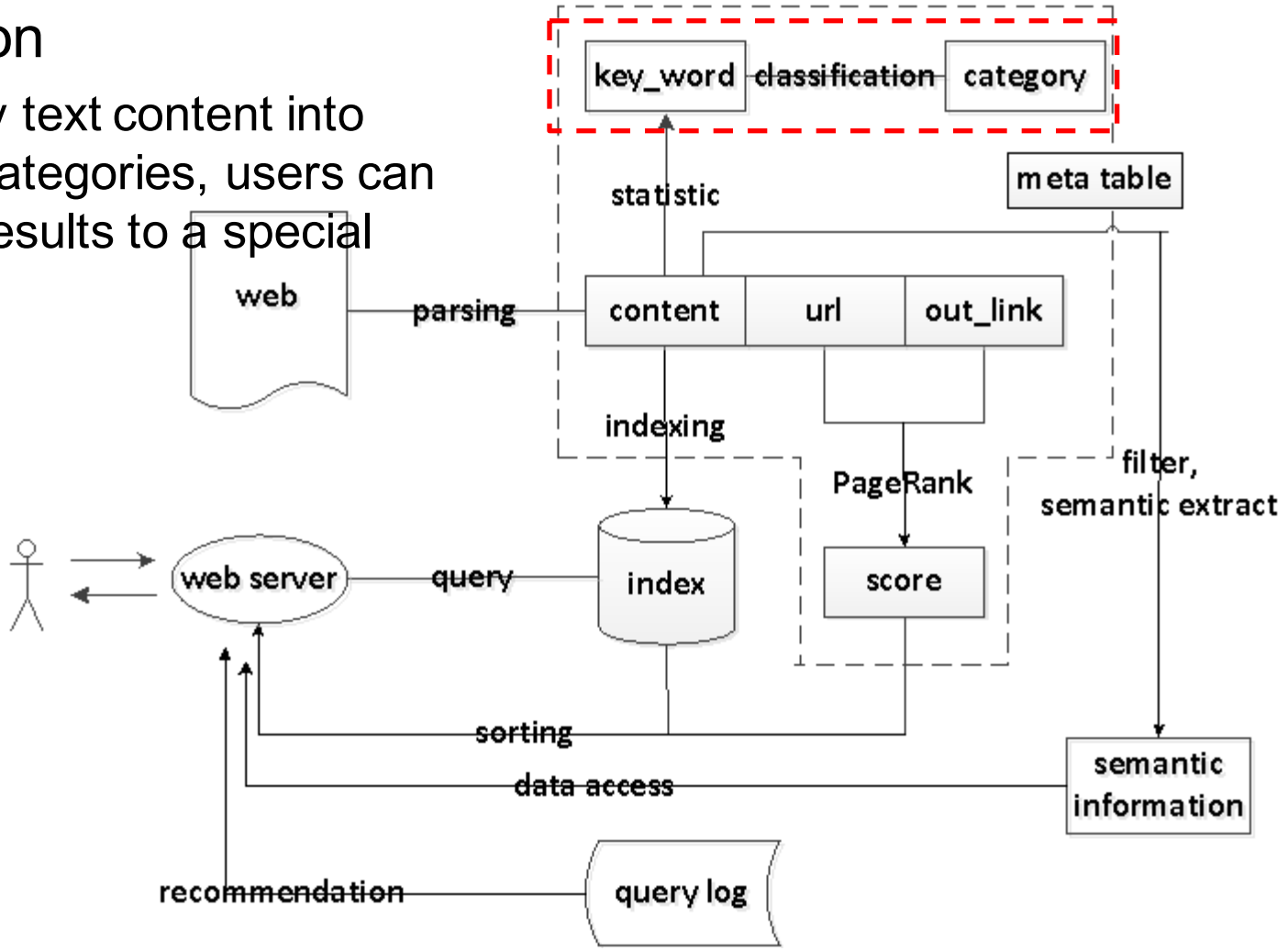
Search Engine: Statistic counting

- Statistic counting
  - To calculate the word frequency to extract the key word which represent the features of the page



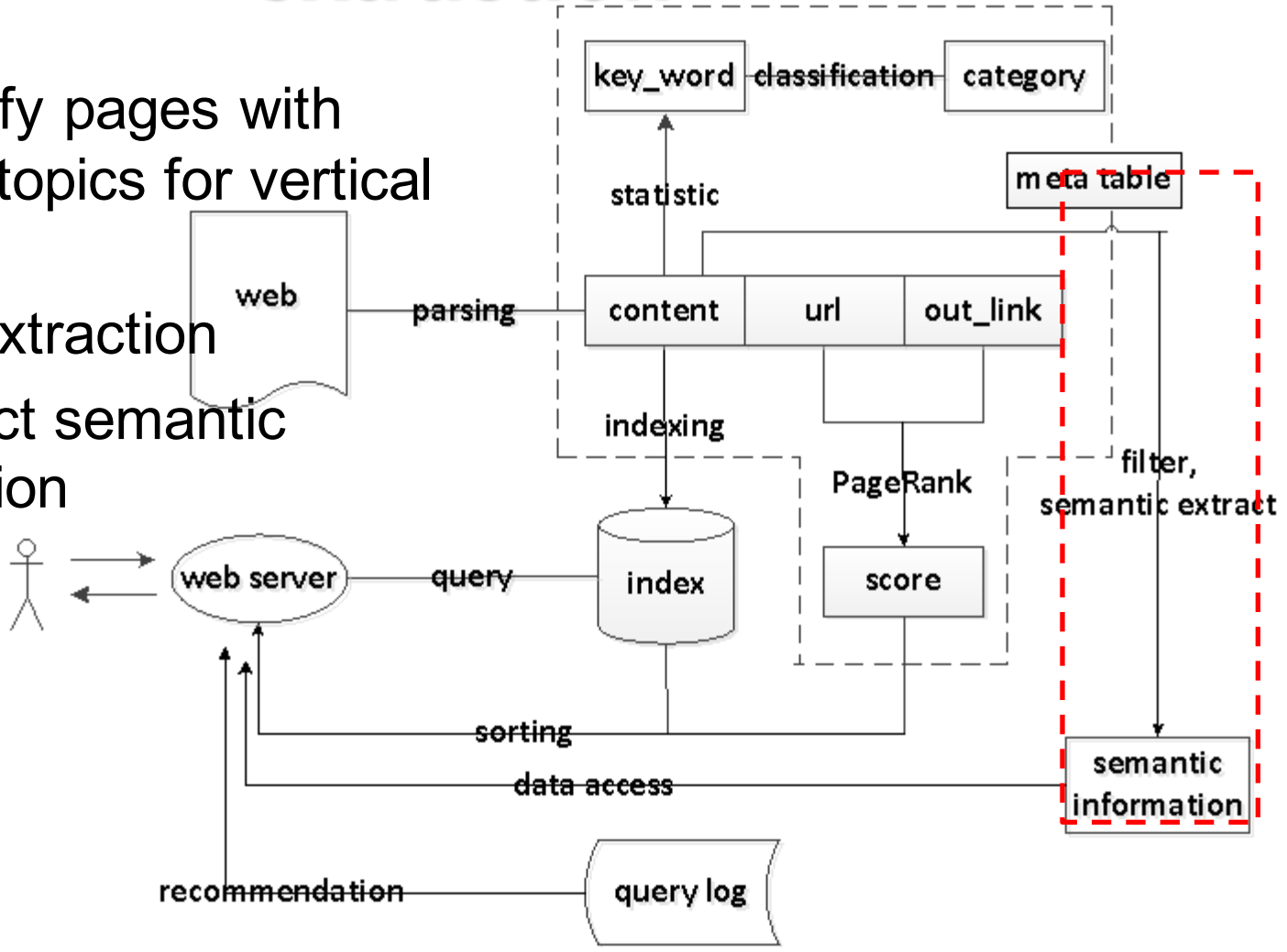
Search Engine: Classification

- Classification
  - To classify text content into different categories, users can filter the results to a special category



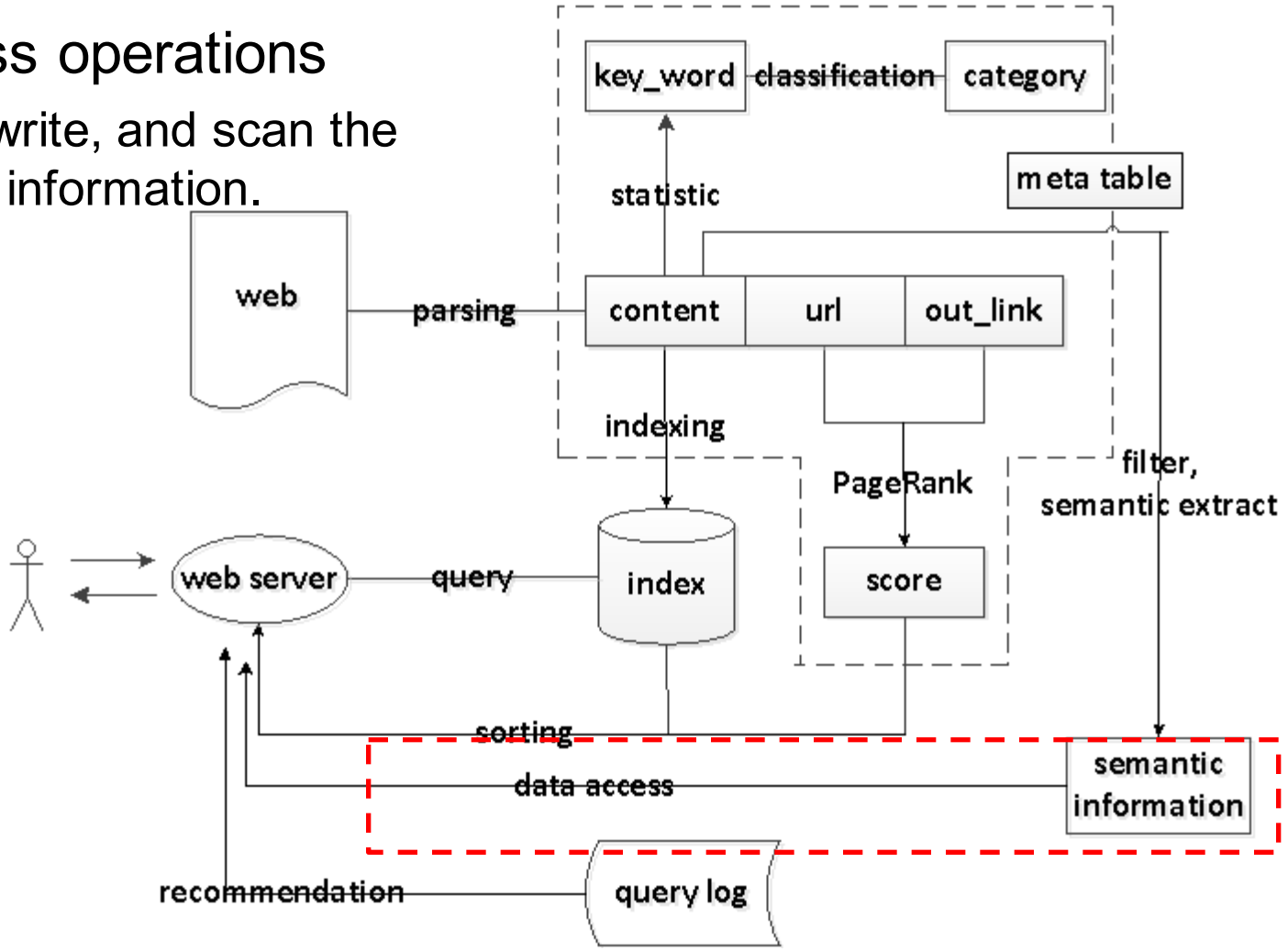
Search Engine: Filter & Semantic extraction

- Filter
  - To identify pages with specific topics for vertical search
- Semantic extraction
  - To extract semantic information



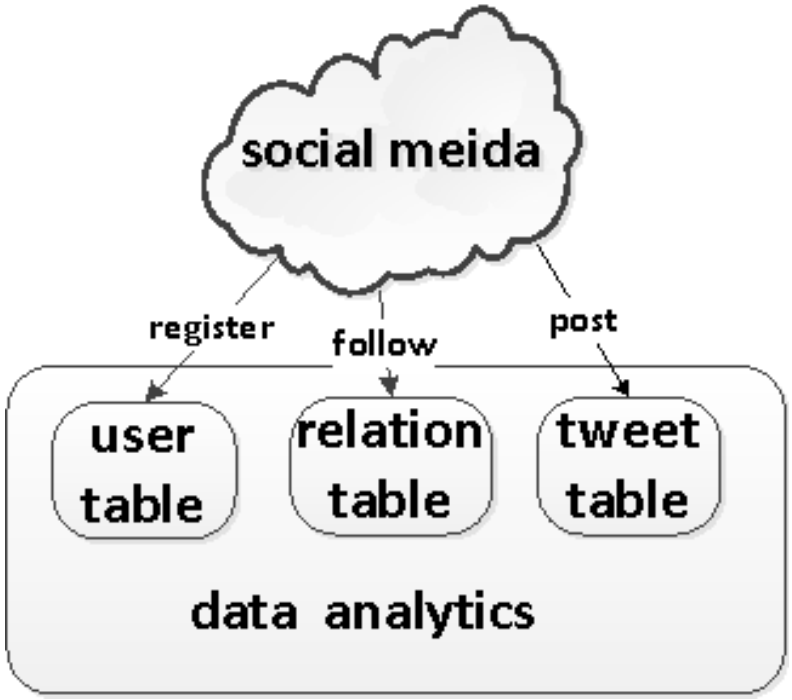
Search Engine: Data access

- Data access operations
  - To read, write, and scan the semantic information.



Social network

- Data sets
  - User table
  - Relation table
  - Article/tweet table
- Workloads
  - Offline analytics



Social network: Data schema

User table

attribute	description
user_id	the id of the user
sex	the sex of the user
age	the age of the user
education	the situation of education
tag	the terms showing characteristics of the user

Relation table

attribute	description
user_id	the id of the user
follow_user_id	the user id who is followed

Tweet table

attribute	description
tweet_id	the id of the tweet
content	the content of the tweet
user_id	the id of user who own the tweet
review_number	the number of review
transmit_number	the number of transmitting
time	the publish time of the tweet

Social network: Workloads

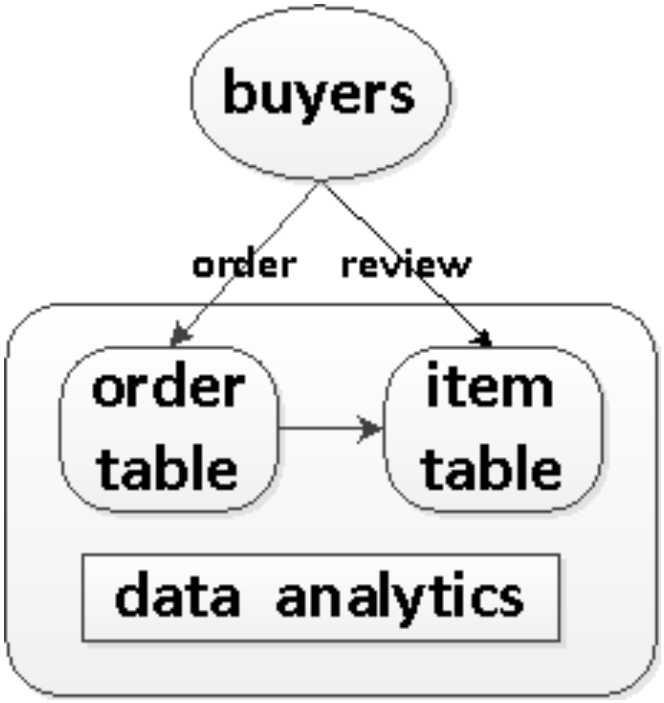
- Hot review topics
  - To select the top N tweets by the number of review
- Hot transmit topics
  - To select the tweets which are transmitted more than N times.
- Active users
  - To select the top N person who posted the largest number of tweets.
- Leaders of opinion
  - To select top ones whose number of review and transmit are both larger than N.

Social network: Workloads

- Topic classification
  - To classify the tweets to certain topics.
- Sentiment classification
  - To classify the tweets to negative or positive according to the sentiment.
- Friend recommendation
  - To recommend friends to person according the relational graph.
- Community detection
  - To detect clusters or communities in large social networks.
- Breadth first search
  - To sort persons according to their distance.

E-commerce

- Data sets
  - Order table
  - Item table
- Workloads
  - Offline analytics



E-commerce: Data schema

Order table

attribute	description
order_id	the id of the order
buyer_id	the id of person who own the order
time	the time of the order occurred

Item table

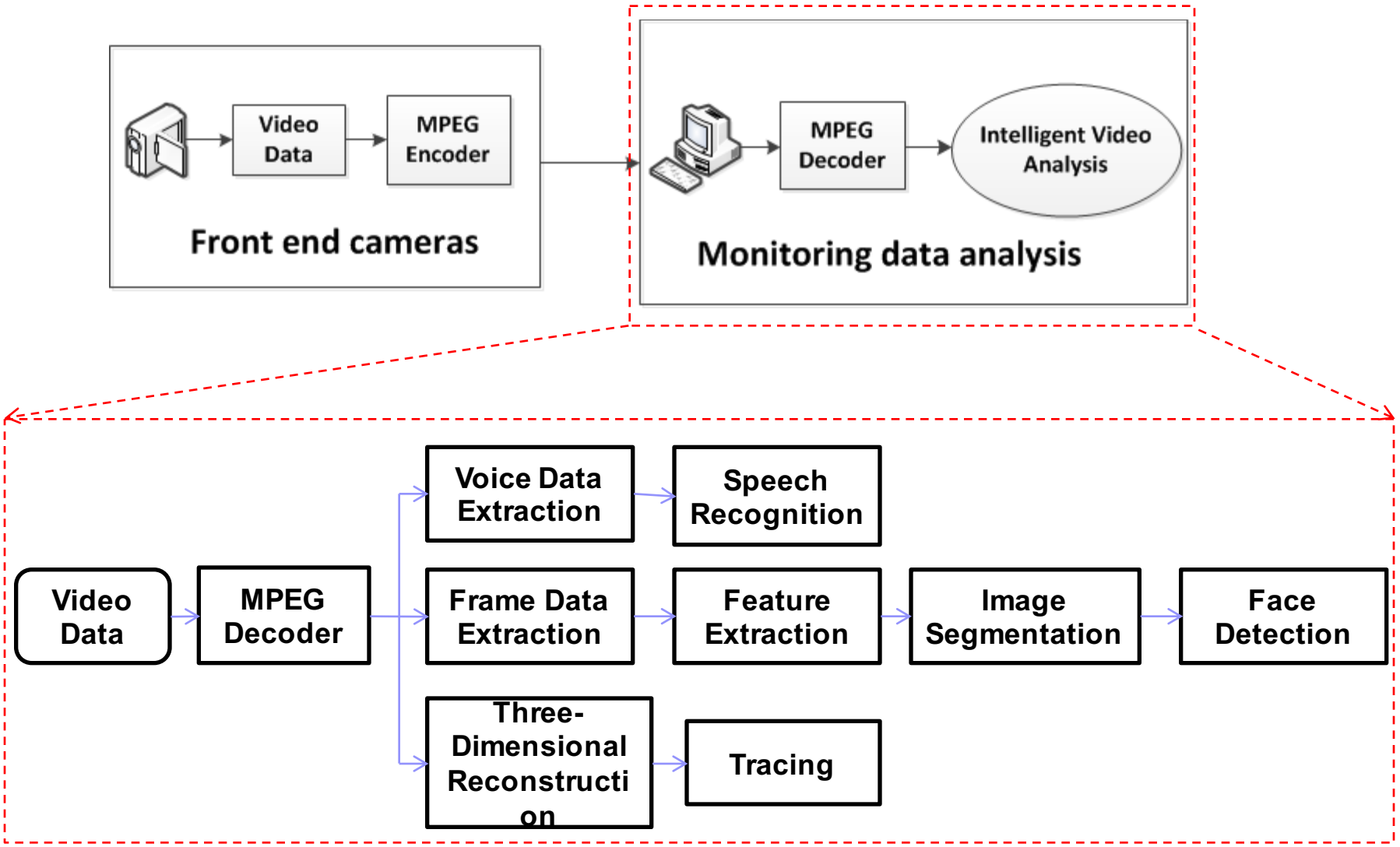
attribute	description
item_id	the id of the item
order_id	the id of order which the item belongs to
goods_id	the id of goods
goods_number	the number of goods
price	the price of goods
amount	the total assumption of the item
score	the score the buyer gave
review	the text commence the buyer gave

E-commerce: Workloads

- Select query
  - To find the items whose sales amount is over 100 in a single order.
- Aggregation query
  - To count the sales number of each goods.
- Join query
  - To count the number of each goods that each buyer purchased between certain period of time.
- Recommendation
  - To predict the preferences of the buyers and recommend goods.
- Sensitive classification
  - To Identify positive or negative review.
- Basic data operation
  - To unit of operation of the data

The workloads of select, aggregation, and join are similar as queries used in A. Pavlo' s SIGMOD'09 paper *,but are specified in the e-commerce environment*

Multimedia


Video Data
MPEG Encoder
Front end cameras
MPEG Decoder
Intelligent Video Analysis
Monitoring data analysis
Video Data
MPEG Decoder
Voice Data Extraction
Speech Recognition
Frame Data Extraction
Feature Extraction
Image Segmentation
Face Detection
Three-Dimensional Reconstruction
Tracing

Multimedia: Workloads

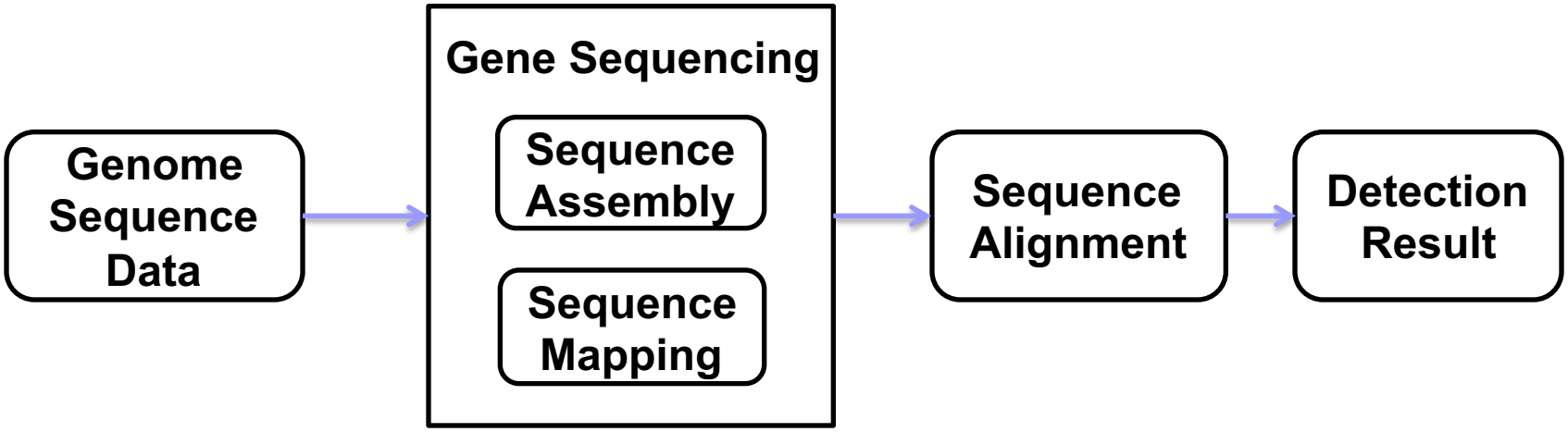
- MPEG Decoder.
  - To decode video streams using MPEG-2 standard.
- Feature extraction
  - For a given video frame, to extract features which are invariant to scale, noise, and illumination.
- Speech Recognition.
  - For a given audio file, to recognize the content of the file and find whether exists sensitive words.

Multimedia: Workloads

- Ray Tracing.
  - To render a 2-Dimensional video frame to a 3-Dimensional scene.
- Image Segmentation.
  - To segment the input video frame according to color, intensity, and texture, and extract concerned regions.
- Face Detection.
  - To detect whether face exists in the input data, if exists, then extract the face.
- Deep Learning.
  - To classify the input images into different categories, and then detect human face.

Bioinformatics

- **Sequence assembly.**
  - To assemble scattered and repetitive DNA fragments to original long sequence.
- **Sequence alignment.**
  - To align assembled DNA sequence to known sequences in the database, and detect disease.



Summary：Real data sets

No.	data sets	data set description[1]	scalable data set
1	Wikipedia Entries	4,300,000 English articles (unstructured text)	Text Generator of BDGS
2	Amazon Movie Reviews	7,911,684 reviews (semi-structured text)	Text Generator of BDGS
3	Google Web Graph	875713 nodes, 5105039 edges (unstructured graph)	Graph Generator of BDGS
4	Facebook Social Network	4039 nodes, 88234 edges (unstructured graph)	Graph Generator of BDGS
5	E-commerce Transaction Data	Table 1: 4 columns, 38658 rows. Table 2: 6 columns, 242735 rows (structured table)	Table Generator of BDGS
6	ProfSearch Person Resumés	278956 resumés (semi-structured table)	Table Generator of BDGS
7	ImageNet	ILSVRC2014 DET image dataset (unstructured image)	ongoing development
8	English broadcasting audio files	Sampled at 16 kHz, 16-bit linear sampling (unstructured audio)	ongoing development
9	DVD Input Streams	110 input streams, resolution:704*480 (unstructured video)	ongoing development
10	Image scene	39 image scene description files (unstructured text)	ongoing development
11	Genome sequence data	cfa data format (unstructured text)	4 volumes of data sets
12	Assembly of the human genome	fa data format (unstructured text)	4 volumes of data sets
13	SoGou Data	the corpus and search query data from SoGou Labs (unstructured text)	ongoing development
14	MNIST	handwritten digits database which has 60,000 training examples and 10,000 test examples (unstructured image)	ongoing development
15	MovieLens Dataset	User's score data for movies, which has 9,518,231 training examples and 386,835 test examples(semi-structured text)	ongoing development

Summary：Search Engine

ID	Implementation	Description	Data set	Software stack
W1-1	Grep	String searching used to parser web pages	Wikipedia data	Hadoop, MPI, Spark, Flink
W1-2	WordCount	Counting the word frequency to do statistic	Wikipedia Data	MPI, Spark, Hadoop, Flink
W1-4	Index	Indexing web pages for searching	Wikipedia data	MPI, Spark, Hadoop
W1-5	PageRank	Computing the importance of the page	Google Web Graph	MPI, Spark, Hadoop, Flink
W1-6-1	Nutch Server	Providing online search services	Sogou Data	Nutch
W1-6-2	Search	Real-time search based on Lucene	Search Data	JStorm
W1-7	Sort	Ordering the data	Wikipedia data	MPI, Spark, Hadoop
W1-11-1	Read	Read operation of data access	Personal Resumes	HBase, Mysql
W1-11-2	Write	Write operators of data access	Personal Resumes	HBase, Mysql
W1-11-3	Scan	Scan operators of data access	Personal Resumes	HBase, Mysql

Various implementation

Summary：Social network

ID	Implementation	Description	Data set	Software stack
W2-1	Rolling Top Words	RollingTopNWord algorithm which used to recommend hot topic	Random Generate	JStorm, Spark Streaming
W2-8-1	CC	Community detection using Connect Component algorithm	Facebook Social Network	MPI, Spark, Hadoop, GraphX, GraphLab, Flink Gelly
W2-8-2	Kmeans	Community detection using Kmeans algorithm	Facebook Social Network	MPI, Spark, Hadoop, Flink, Spark Streaming
W2-8-3	Label Propagation	Label propagation algorithm for community detection in graphs	Facebook Social Network	GraphX, GraphLab, Flink Gelly
W2-8-4	Triangle Count	Triangle count algorithm for community detection in graphs	Facebook Social Network	GraphX, GraphLab, Flink Gelly
W2-9	BFS	Breadth first search	synthetic graph	MPI, GraphX, GraphLab, Flink Gelly

Summary：E-commerce

ID	Implementation	Description	Data set	Software stack
W3-1	Select query	Find the items of which the sales amount is over 100 in a single order	E-commence Transaction	Hive, Shark, Impala
W3-2	Aggregation query	Count the sales number of each goods	E-commence Transaction	Hive, Shark, Impala
W3-3	Join query	Count the number of each goods that each buyer purchased between certain period of time	E-commence Transaction	Hive, Shark, Impala
W3-4	CF	Recommendation using Collaborative Filtering algorithm	Amazon Movie Reviews	Hadoop, Spark, MPI, JStorm
W3-5	Native Bayes	Sensitive classification using Native Bayes algorithm	Amazon Movie Reviews	Hadoop, Spark, MPI
W3-6-1	Project	Basic operator	E-commerce Transaction	Hive, Shark, Impala
W3-6-2	Filter	Basic operator	E-commerce Transaction	Hive, Shark, Impala
W3-6-3	Cross Product	Basic operator	E-commerce Transaction	Hive, Shark, Impala
W3-6-4	OrderBy	Basic operator	E-commerce Transaction	Hive, Shark, Impala
W3-6-5	Union	Basic operator	E-commerce Transaction	Hive, Shark, Impala
W3-6-6	Difference	Basic operator	E-commerce Transaction	Hive, Shark, Impala
W3-6-7	Aggregation	Basic operator	E-commerce Transaction	Hive, Shark, Impala

Summary：Multimedia

ID	Implementation	Description	Data Set	Software Stack
W4-1	BasicMPEG [43]	MPEG2 decode/encode	DVD Input Streams	Libc
W4-2-1	SIFT [48]	Detect and describe local features in input images	ImageNet	MPI
W4-2-2	DBN [48]	Implementation of Deep Belief Networks	MNIST	MPI
W4-3	Speech Recognition [6]	Translate spoken words into text	English broadcasting audio files	MPI
W4-4	Ray Tracing [60]	Generating an 3D image by tracing light	Image scene	MPI
W4-5	Image Segmentation [29]	Partitioning an image into multiple segments	ImageNet	MPI
W4-6	Face Detection [61]	Detecting face in an image	ImageNet	MPI

Summary：Bioinformatics

ID	Implementation	Description	Data Set	Software Stack
W5-1	SAND	Sequence assembly implementations which merge genome fragments to get the original genome sequence	Genome sequence data	Work Queue
W5-2	BLAST	Sequence alignment implementations which identify the similarity between target sequence with sequence in database	Assembly of the human genome data	MPI


Any
Questions