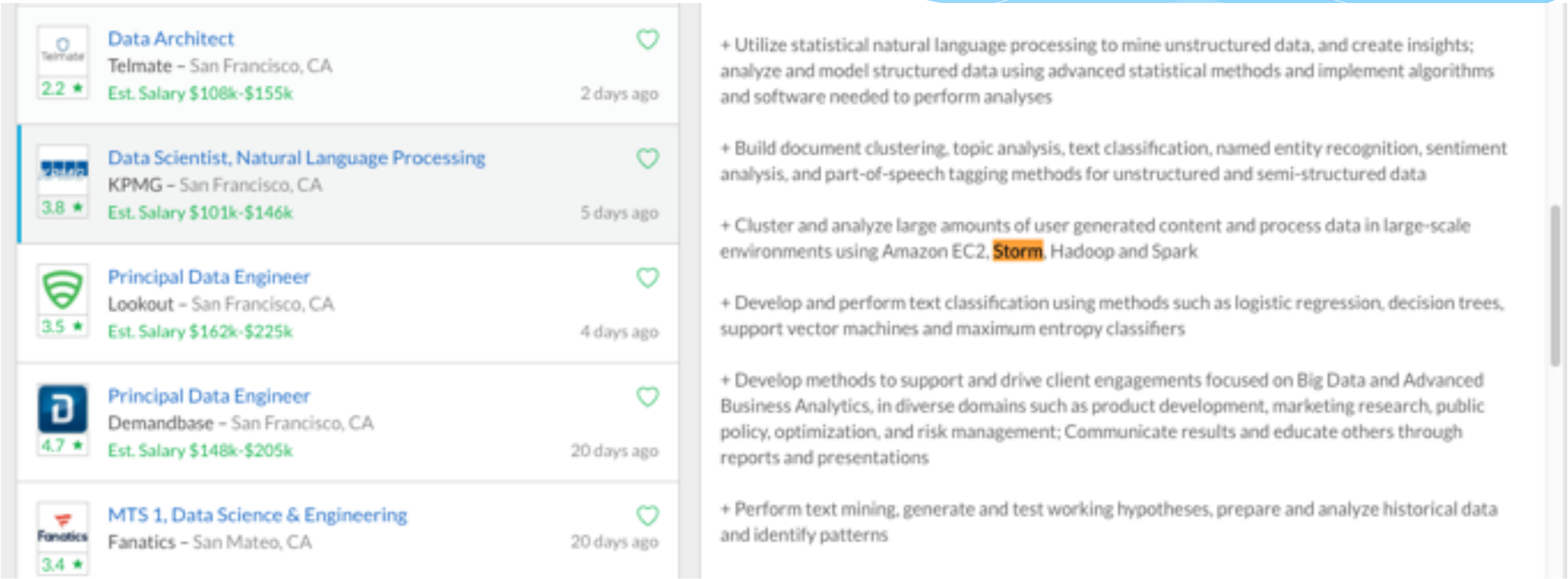
405 MAS- Distributed Stream Processing

Christian Gao

INCENTIVES



Telmate
2.2 ★
Data Architect
Telmate – San Francisco, CA
Est. Salary $108k-$155k
2 days ago

3.8 ★
Data Scientist, Natural Language Processing
KPMG – San Francisco, CA
Est. Salary $101k-$146k
5 days ago

3.5 ★
Principal Data Engineer
Lookout – San Francisco, CA
Est. Salary $162k-$225k
4 days ago

4.7 ★
Principal Data Engineer
Demandbase – San Francisco, CA
Est. Salary $148k-$205k
20 days ago

Fanatics
3.4 ★
MTS 1, Data Science & Engineering
Fanatics – San Mateo, CA
20 days ago

+ Utilize statistical natural language processing to mine unstructured data, and create insights; analyze and model structured data using advanced statistical methods and implement algorithms and software needed to perform analyses

+ Build document clustering, topic analysis, text classification, named entity recognition, sentiment analysis, and part-of-speech tagging methods for unstructured and semi-structured data

+ Cluster and analyze large amounts of user generated content and process data in large-scale environments using Amazon EC2, Storm, Hadoop and Spark

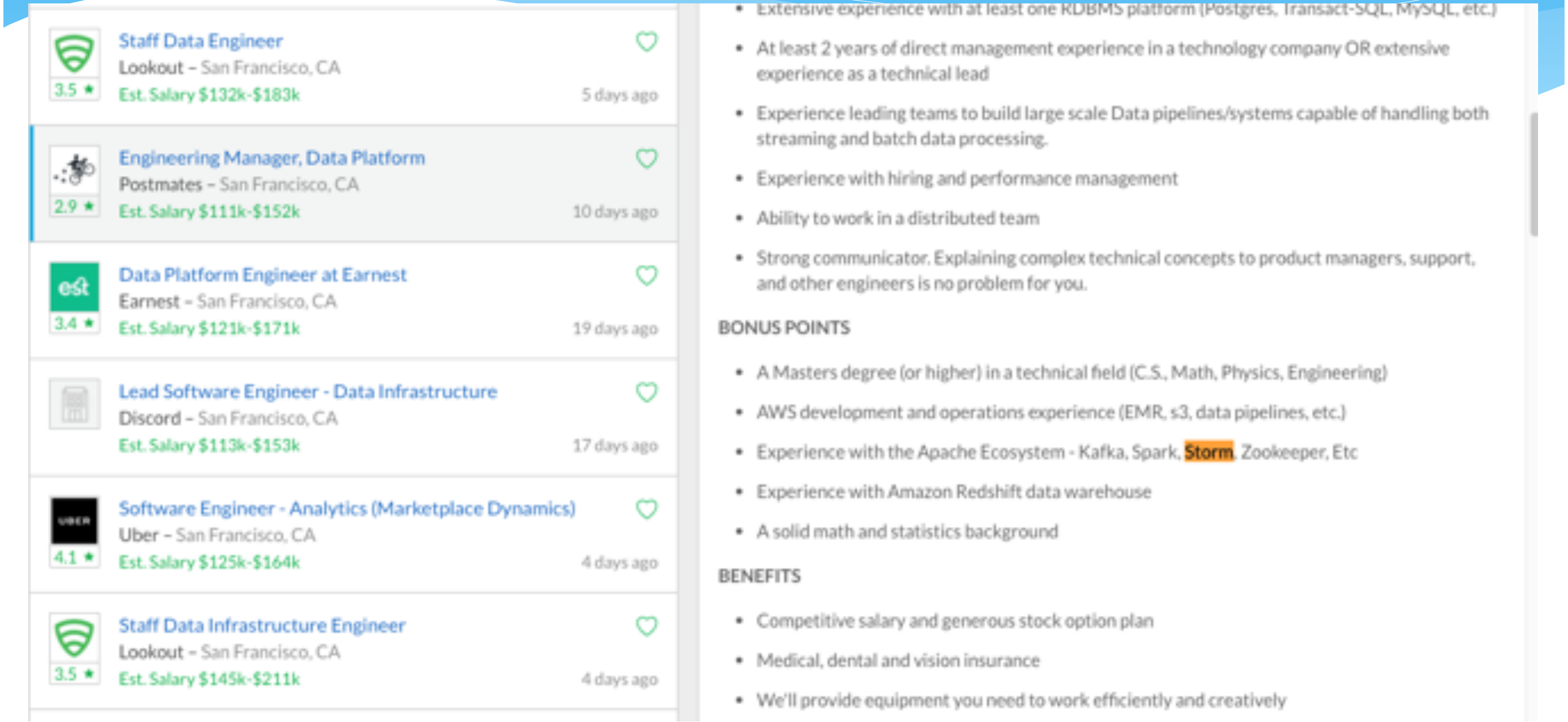
+ Develop and perform text classification using methods such as logistic regression, decision trees, support vector machines and maximum entropy classifiers

+ Develop methods to support and drive client engagements focused on Big Data and Advanced Business Analytics, in diverse domains such as product development, marketing research, public policy, optimization, and risk management; Communicate results and educate others through reports and presentations

+ Perform text mining, generate and test working hypotheses, prepare and analyze historical data and identify patterns

Data Application Lab

INCENTIVES cont. Manager



Staff Data Engineer
Lookout – San Francisco, CA
Est. Salary $132k-$183k
3.5 ★
5 days ago

Engineering Manager, Data Platform
Postmates – San Francisco, CA
Est. Salary $111k-$152k
2.9 ★
10 days ago

Data Platform Engineer at Earnest
Earnest – San Francisco, CA
Est. Salary $121k-$171k
3.4 ★
19 days ago

Lead Software Engineer - Data Infrastructure
Discord – San Francisco, CA
Est. Salary $113k-$153k
17 days ago

Software Engineer - Analytics (Marketplace Dynamics)
Uber – San Francisco, CA
Est. Salary $125k-$164k
4.1 ★
4 days ago

Staff Data Infrastructure Engineer
Lookout – San Francisco, CA
Est. Salary $145k-$211k
3.5 ★
4 days ago

- Extensive experience with at least one RDBMS platform (Postgres, Transact-SQL, MySQL, etc.)
- At least 2 years of direct management experience in a technology company OR extensive experience as a technical lead
- Experience leading teams to build large scale Data pipelines/systems capable of handling both streaming and batch data processing.
- Experience with hiring and performance management
- Ability to work in a distributed team
- Strong communicator. Explaining complex technical concepts to product managers, support, and other engineers is no problem for you.

BONUS POINTS

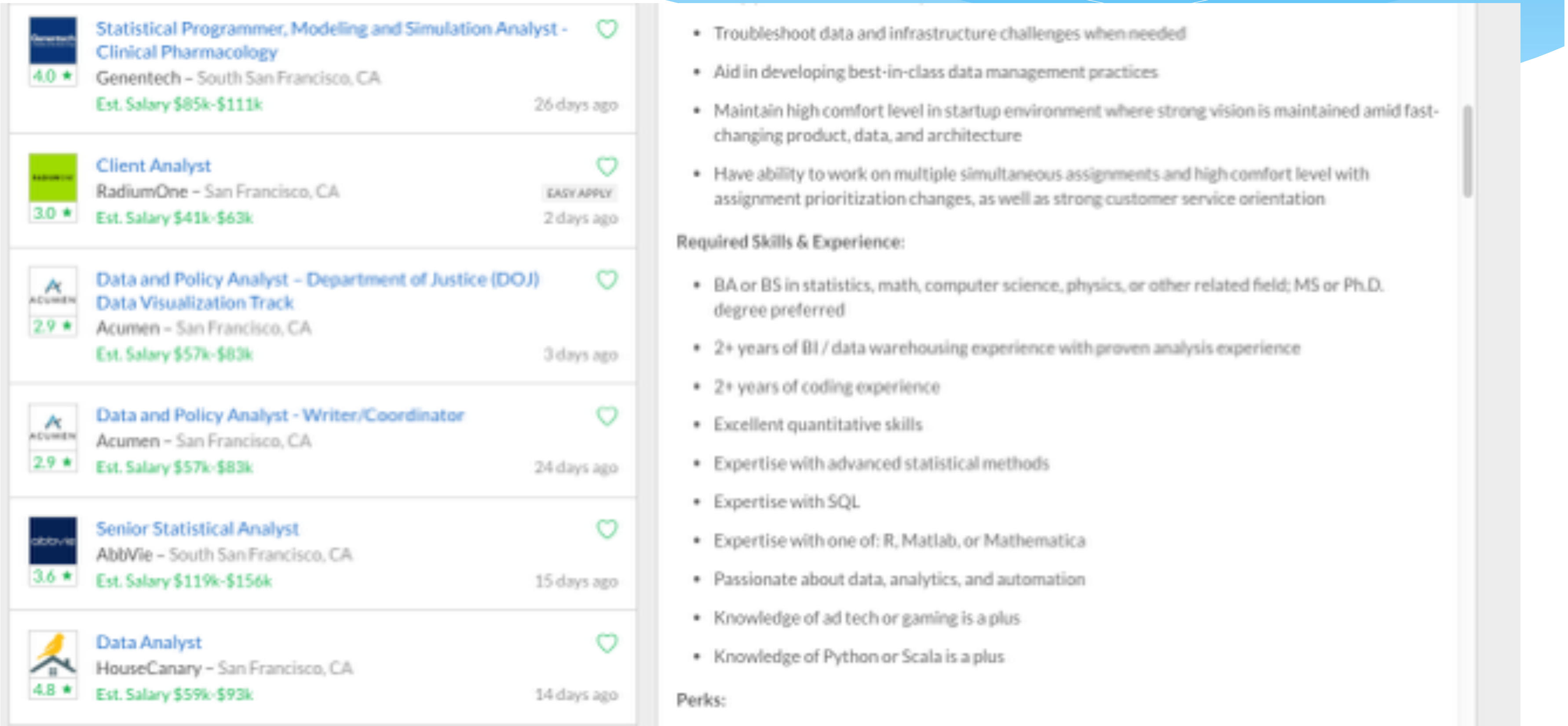
- A Masters degree (or higher) in a technical field (C.S., Math, Physics, Engineering)
- AWS development and operations experience (EMR, s3, data pipelines, etc.)
- Experience with the Apache Ecosystem - Kafka, Spark, Storm, Zookeeper, Etc
- Experience with Amazon Redshift data warehouse
- A solid math and statistics background

BENEFITS

- Competitive salary and generous stock option plan
- Medical, dental and vision insurance
- We'll provide equipment you need to work efficiently and creatively

Data Application Lab

INCENTIVES - NO Distributed Streaming



Statistical Programmer, Modeling and Simulation Analyst - Clinical Pharmacology
4.0 ★
Genentech – South San Francisco, CA
Est. Salary $85k-$111k
26 days ago

Client Analyst
3.0 ★
RadiumOne – San Francisco, CA
EASY APPLY
Est. Salary $41k-$63k
2 days ago

Data and Policy Analyst – Department of Justice (DOJ) Data Visualization Track
2.9 ★
Acumen – San Francisco, CA
Est. Salary $57k-$83k
3 days ago

Data and Policy Analyst - Writer/Coordinator
2.9 ★
Acumen – San Francisco, CA
Est. Salary $57k-$83k
24 days ago

Senior Statistical Analyst
3.6 ★
AbbVie – South San Francisco, CA
Est. Salary $119k-$156k
15 days ago

Data Analyst
4.8 ★
HouseCanary – San Francisco, CA
Est. Salary $59k-$93k
14 days ago

- Troubleshoot data and infrastructure challenges when needed
- Aid in developing best-in-class data management practices
- Maintain high comfort level in startup environment where strong vision is maintained amid fast-changing product, data, and architecture
- Have ability to work on multiple simultaneous assignments and high comfort level with assignment prioritization changes, as well as strong customer service orientation

Required Skills & Experience:

- BA or BS in statistics, math, computer science, physics, or other related field; MS or Ph.D. degree preferred
- 2+ years of BI / data warehousing experience with proven analysis experience
- 2+ years of coding experience
- Excellent quantitative skills
- Expertise with advanced statistical methods
- Expertise with SQL
- Expertise with one of: R, Matlab, or Mathematica
- Passionate about data, analytics, and automation
- Knowledge of ad tech or gaming is a plus
- Knowledge of Python or Scala is a plus

Perks:

Data Application Lab

Used By

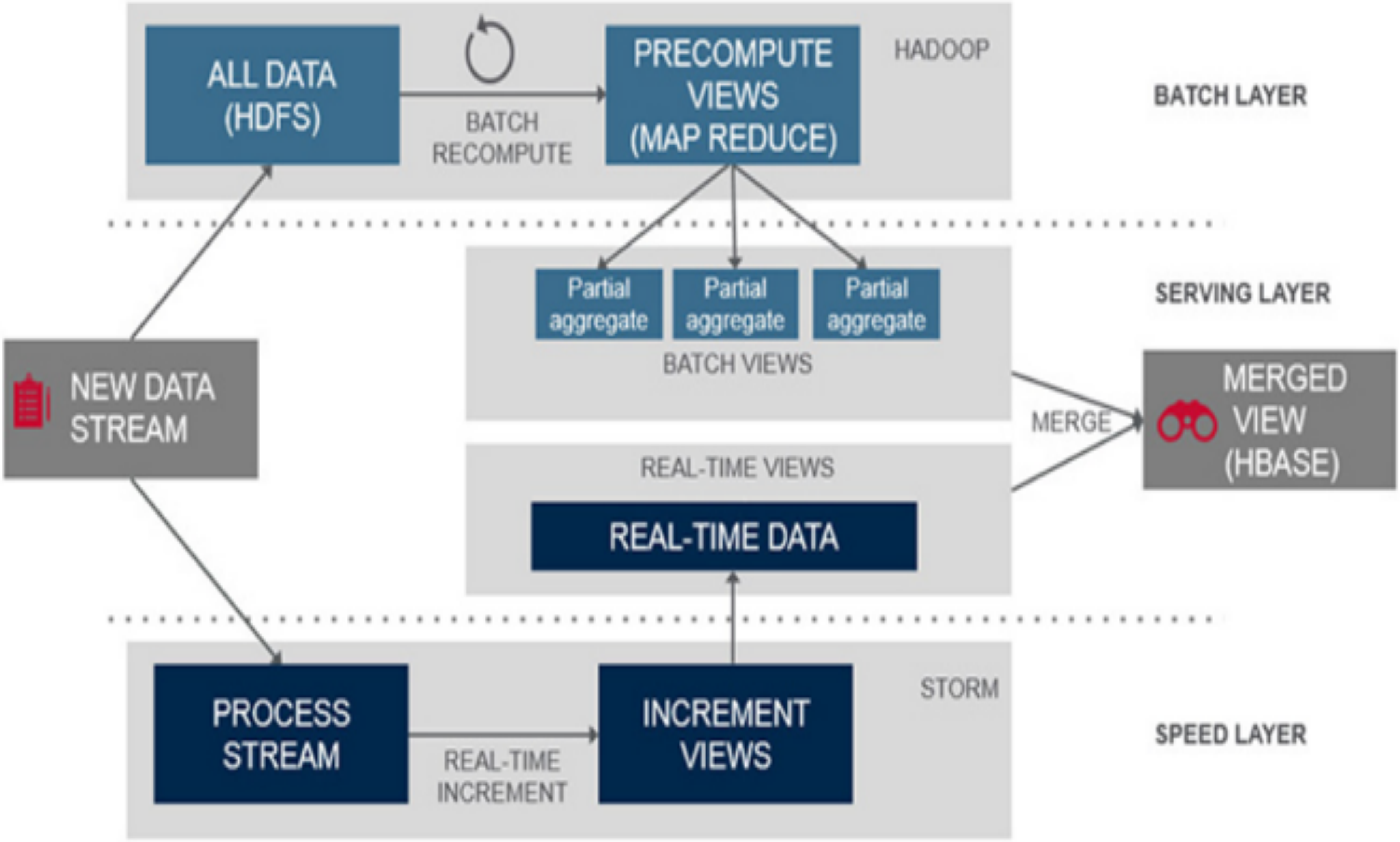
- Spotify
- Groupon
- Twitter
- Yahoo
- WebMD
- TaoBao
- Alibaba
- Baidu
- Yelp

Distributed Computing vs Regular Computing

- Cluster Talking to Cluster
- Multiple Threads
- Less Speed More Throughput
- Multi Thread Link:
  - http://beginnersbook.com/2013/03/multithreading-in-java/

Data Application Lab

Lambda Architecture



Characteristics

- Fast – benchmarked as processing one million 100 byte messages per second per node
- Scalable – with parallel calculations that run across a cluster of machines
- Fault-tolerant – when workers die, Storm will automatically restart them. If a node dies, the worker will be restarted on another node
- Reliable – Storm guarantees that each unit of data (tuple) will be processed at least once or exactly once. Messages are only replayed when there are failures.
- Easy to operate – standard configurations are suitable for production on day one. Once deployed, Storm is easy to operate

Data Application Lab

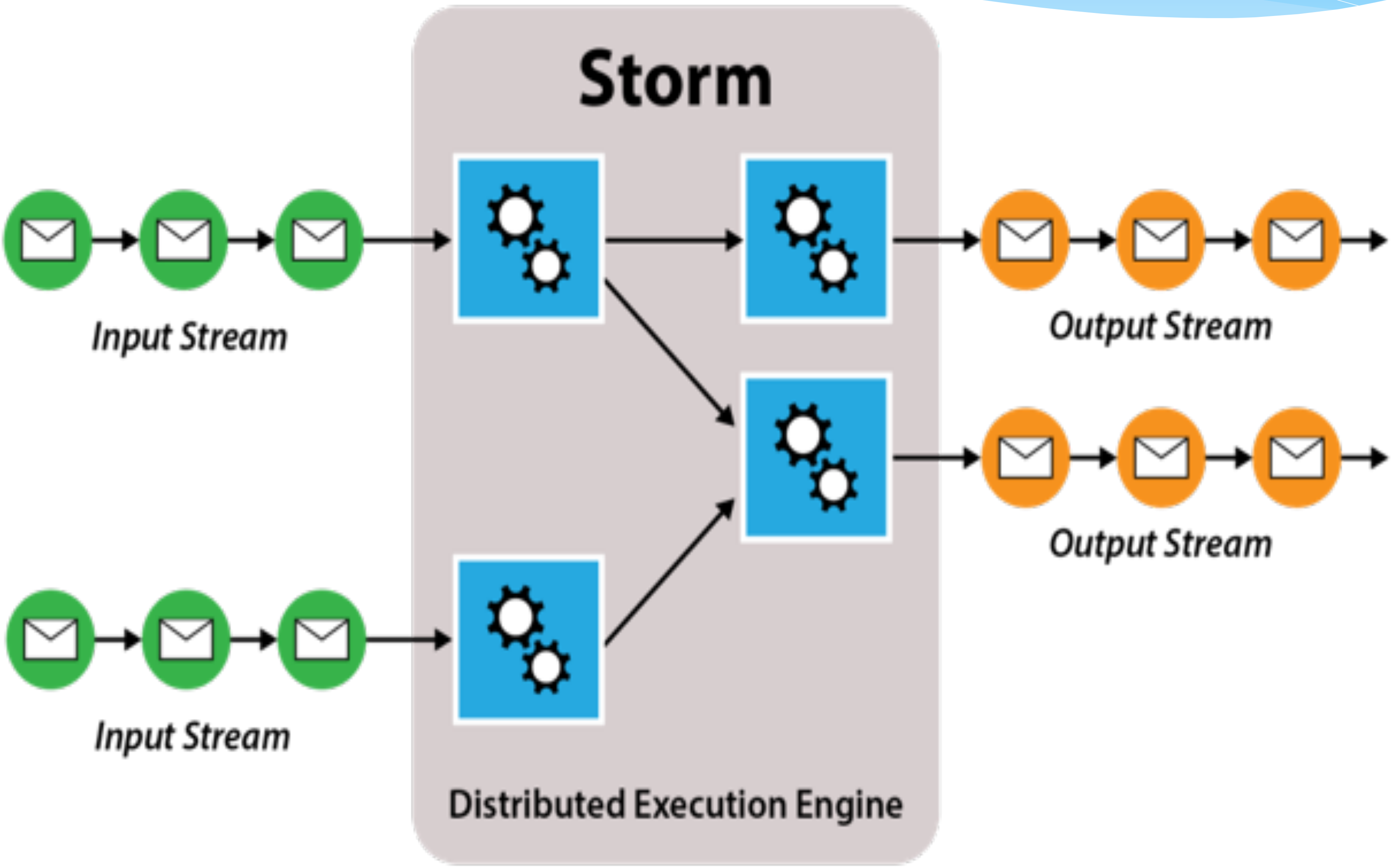
INTERVIEW QUESTION

- Q: Suppose you have … blah blah blah … with conditions … blah blah blah … how would you solve it?

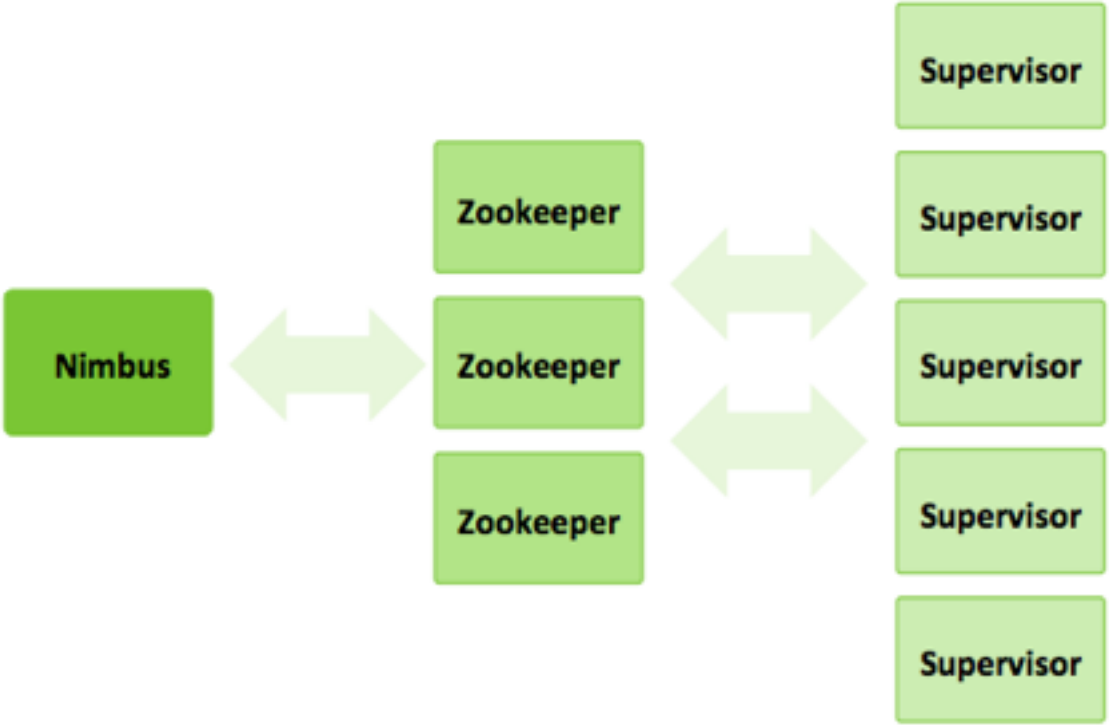
- A: Lambda

- Why: Its impressive to know.

Data Application Lab


Storm
Input Stream
Input Stream
Output Stream
Output Stream
Distributed Execution Engine

Storm Cluster Architecture



Data Application Lab

Storm Components

- Nimbus node (master node, similar to the Hadoop JobTracker):
- - Uploads computations for execution
- - Distributes code across the cluster
- - Launches workers across the cluster
- - Monitors computation and reallocates workers as needed
- - Now at failure cluster will invoke leader election

Data Application Lab

Storm Supervisor

- Supervisor nodes – communicates with Nimbus through Zookeeper, starts and stops workers according to signals from Nimbus
- At a particular time interval, all supervisors will send heartbeats to the nimbus to inform that they are still alive.
- When a supervisor dies and doesn’t send a heartbeat to the nimbus, then the nimbus assigns the tasks to another supervisor.

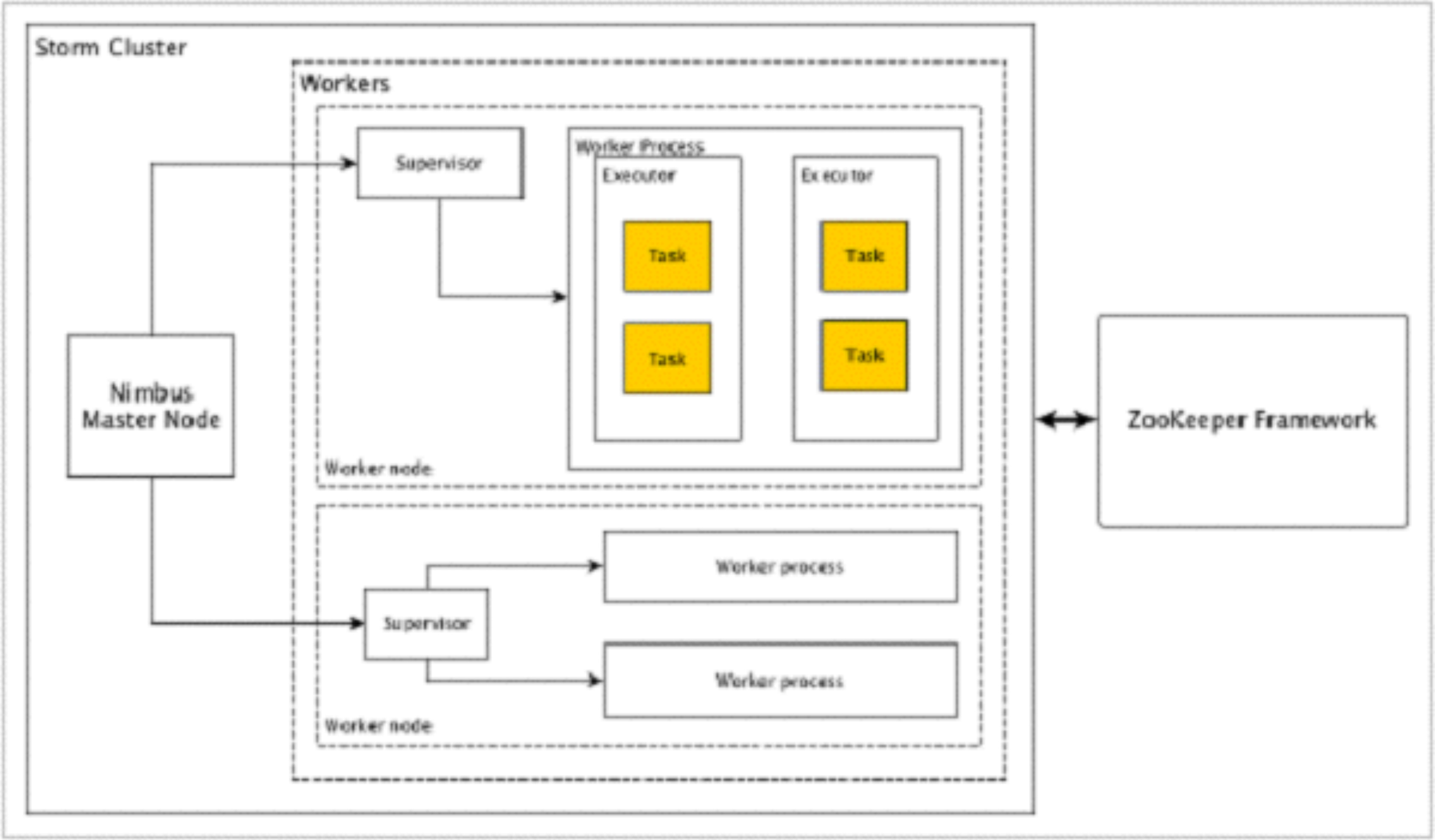
More Components

- ZooKeeper nodes – coordinates the Storm cluster
- - ZooKeeper helps the supervisor to interact with the nimbus.
- - It is responsible to maintain the state of nimbus and supervisor.
- - This is where your storm settings are stored.
- - Usually hosted on 1 or 3 nodes.

Data Application Lab

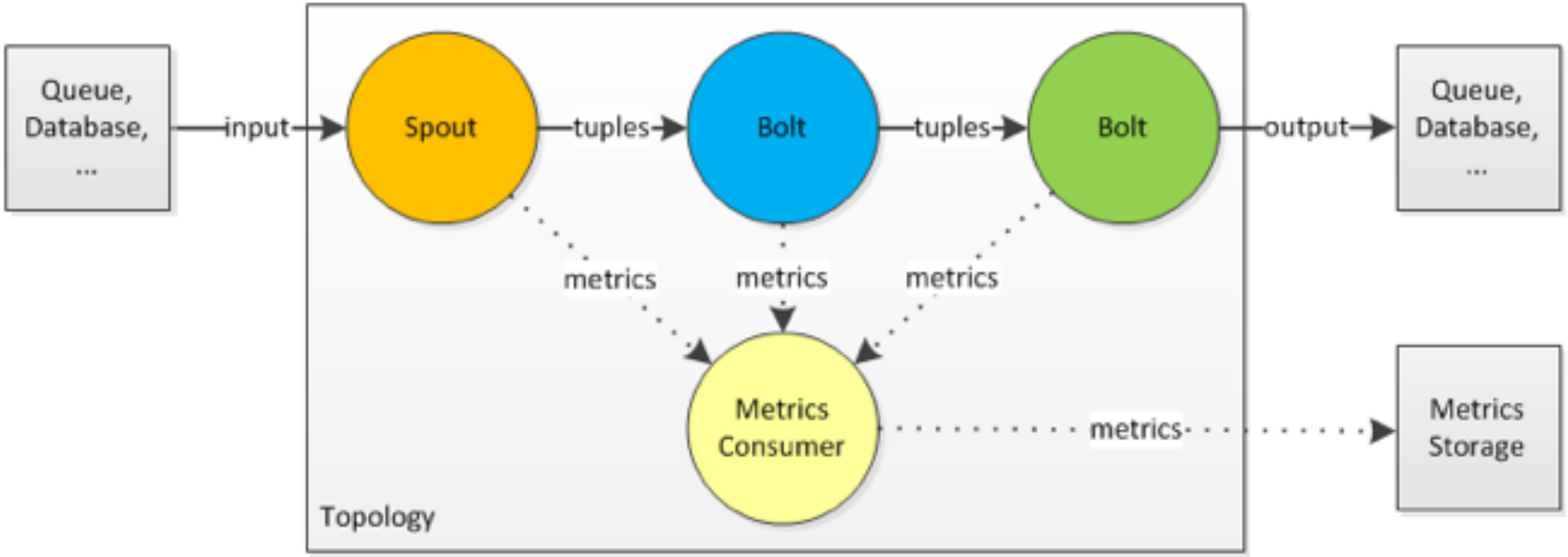
Storm Worker

- Worker- executes tasks related to a specific topology.
- A worker process will not run a task by itself, instead it creates executors.
- Executor- An executor is nothing but a single thread spawn by a worker process.
- Task- A task performs actual data processing. (Ex: Spout or Bolt).


Storm Cluster
Workers
Supervisor
Worker Process
Executor
Executor
Task
Task
Task
Task
Nimbus
Master Node
ZooKeeper Framework
Worker node
Worker process
Supervisor
Worker process
Worker node

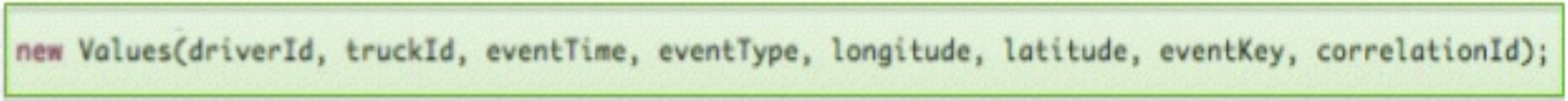
Topology

- Tuples– an ordered list of elements. For example, a “4-tuple” might be (7, 1, 3, 7)
- Streams – an unbounded sequence of tuples.
- Spouts – sources of streams in a computation (e.g. a Twitter API)
- Bolts – process input streams and produce output streams. They can: run functions; filter, aggregate, or join data; or talk to databases.
- Topologies – the overall abstract infrastructure, represented visually as a network of spouts and bolts (as in the following diagram)


Queue,
Database,
...
input
Spout
tuples
Bolt
tuples
Bolt
output
Queue,
Database,
...
metrics
metrics
metrics
Metrics
Consumer
metrics
Metrics
Storage
Topology

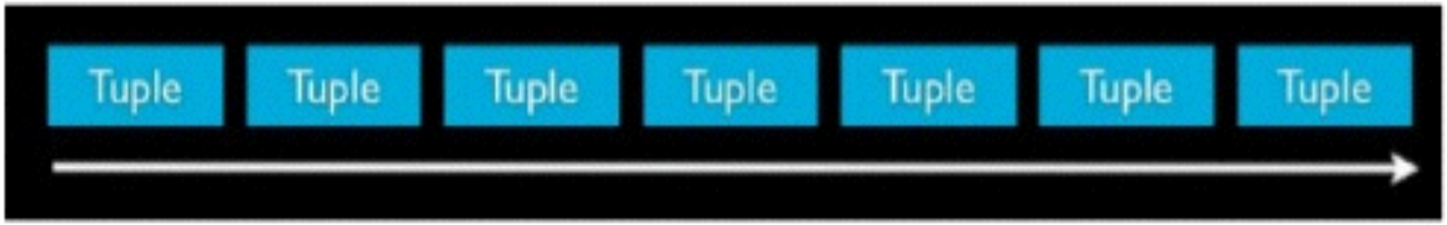
Tuples and Streams

- **What is a Tuple?**
  - Fundamental data structure in Storm. Is a named list of values that can be of any data type.



```
new Values(driverId, truckId, eventTime, eventType, longitude, latitude, eventKey, correlationId);
```

- **What is a Stream?**
  - An unbounded sequences of tuples.
  - Core abstraction in Storm and are what you “process” in Storm

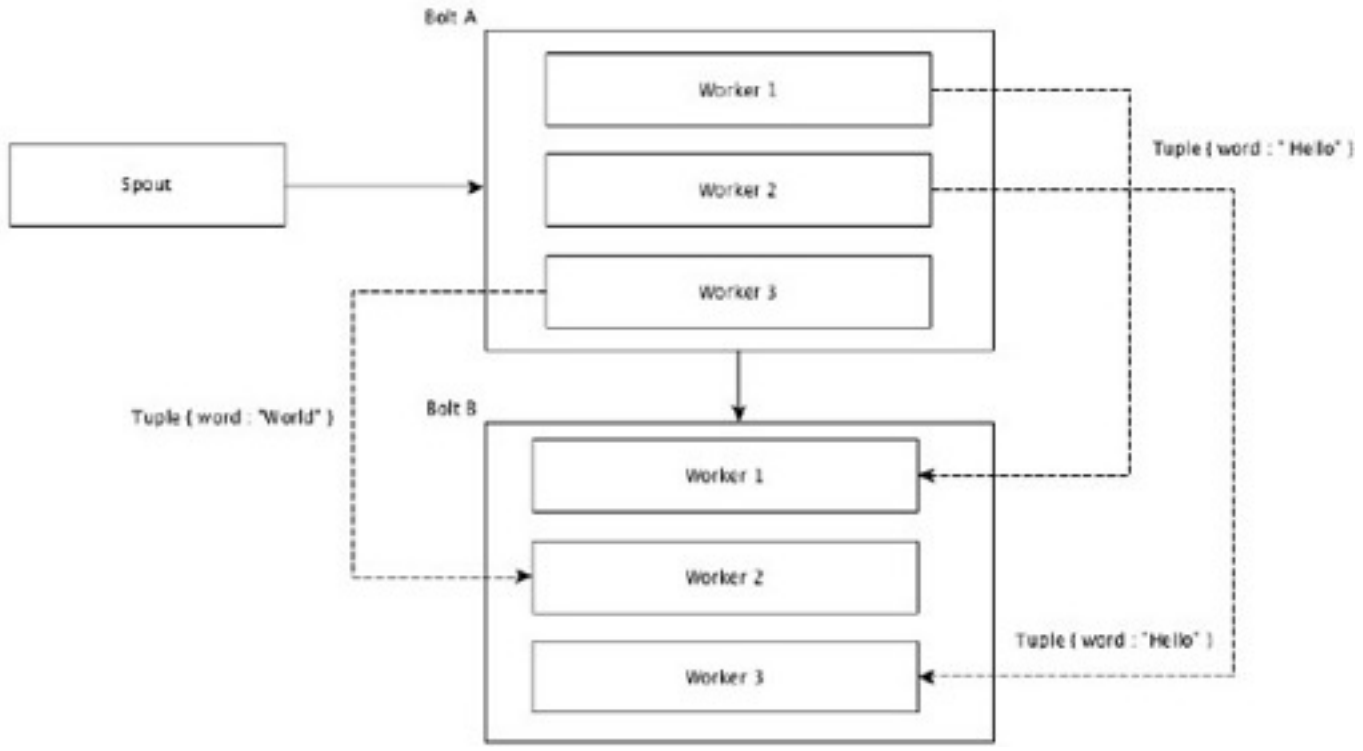


Demo Time

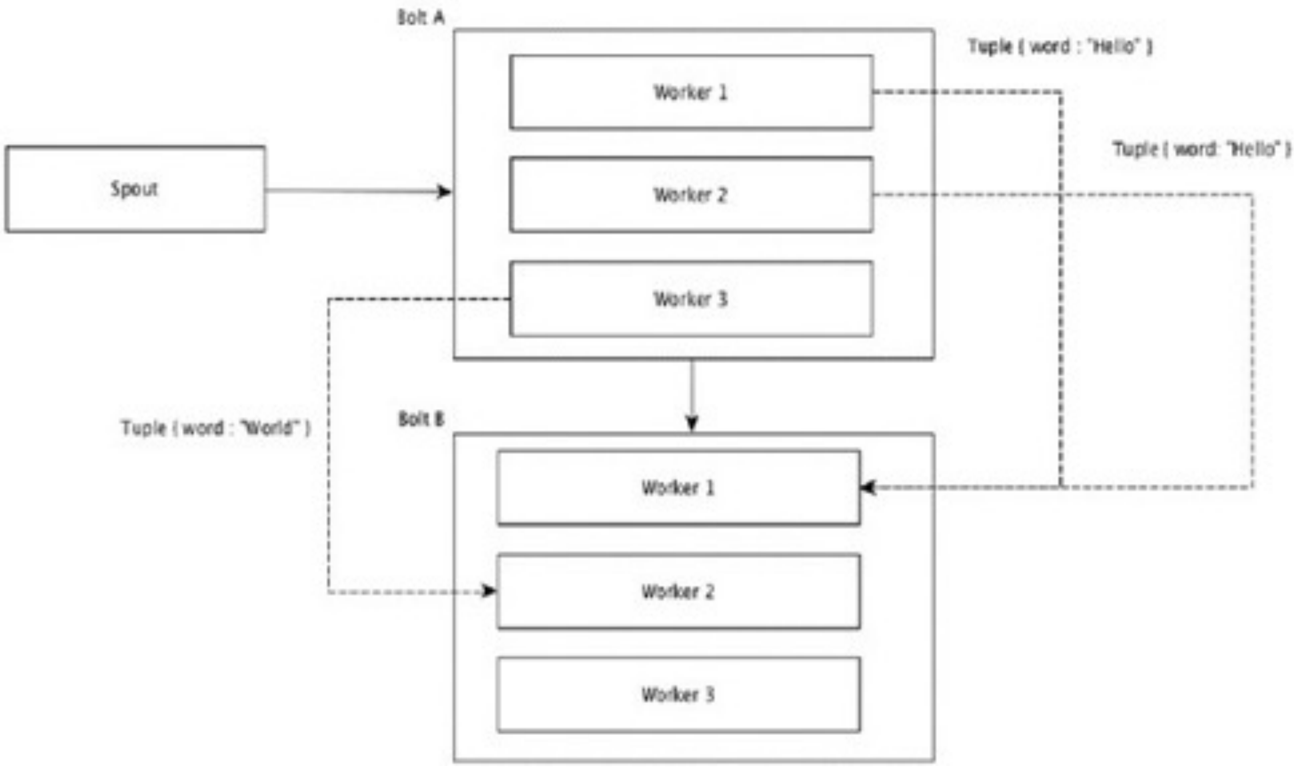
- Cluster Mode: storm jar cluster-topology-jar class
- Local Mode: java -cp local-topology-jar class
- -The Jars and code for cluster and for local are different. For example the jar for local contains code for storm while the jar for the cluster excludes storm.

Data Application Lab

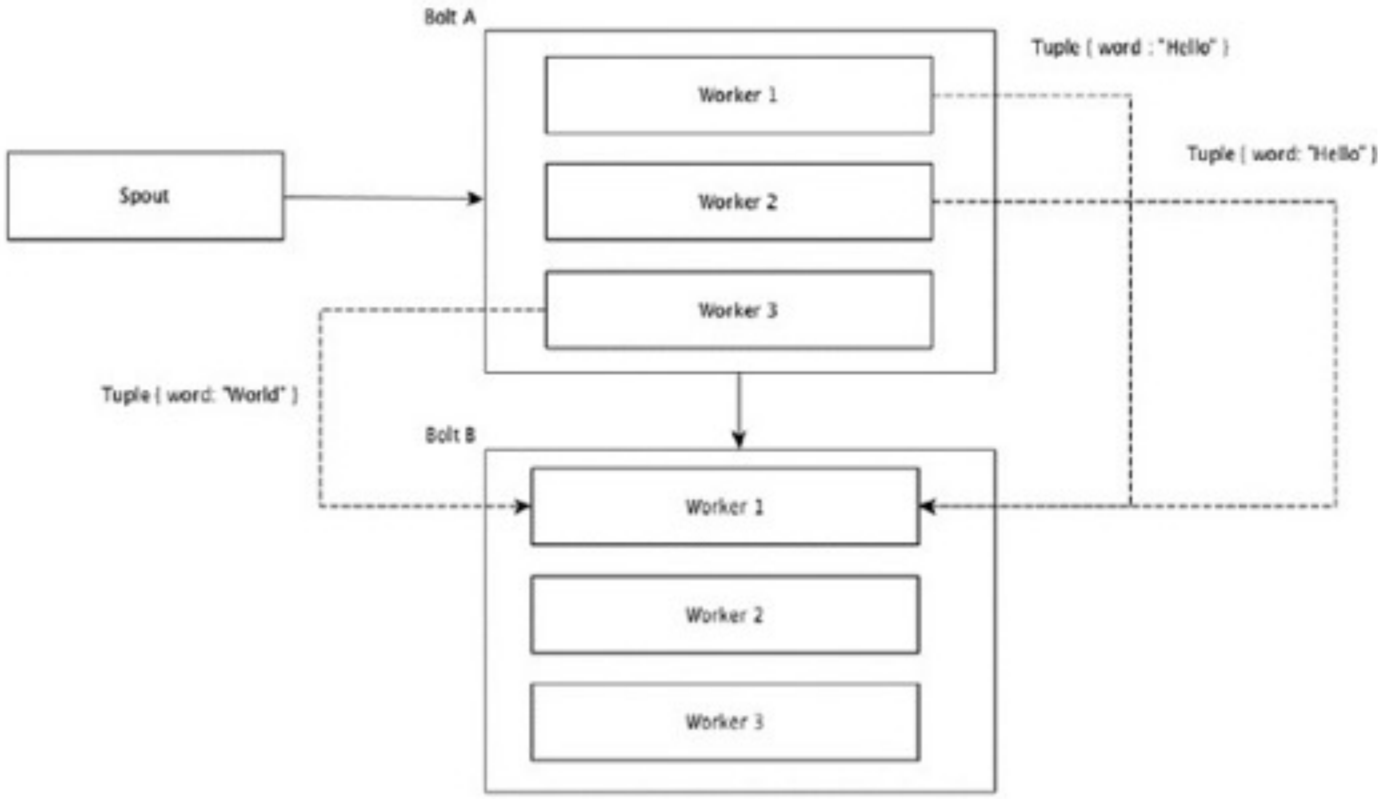
- Shuffle Grouping- an equal number of tuples is distributed randomly across all of the workers.



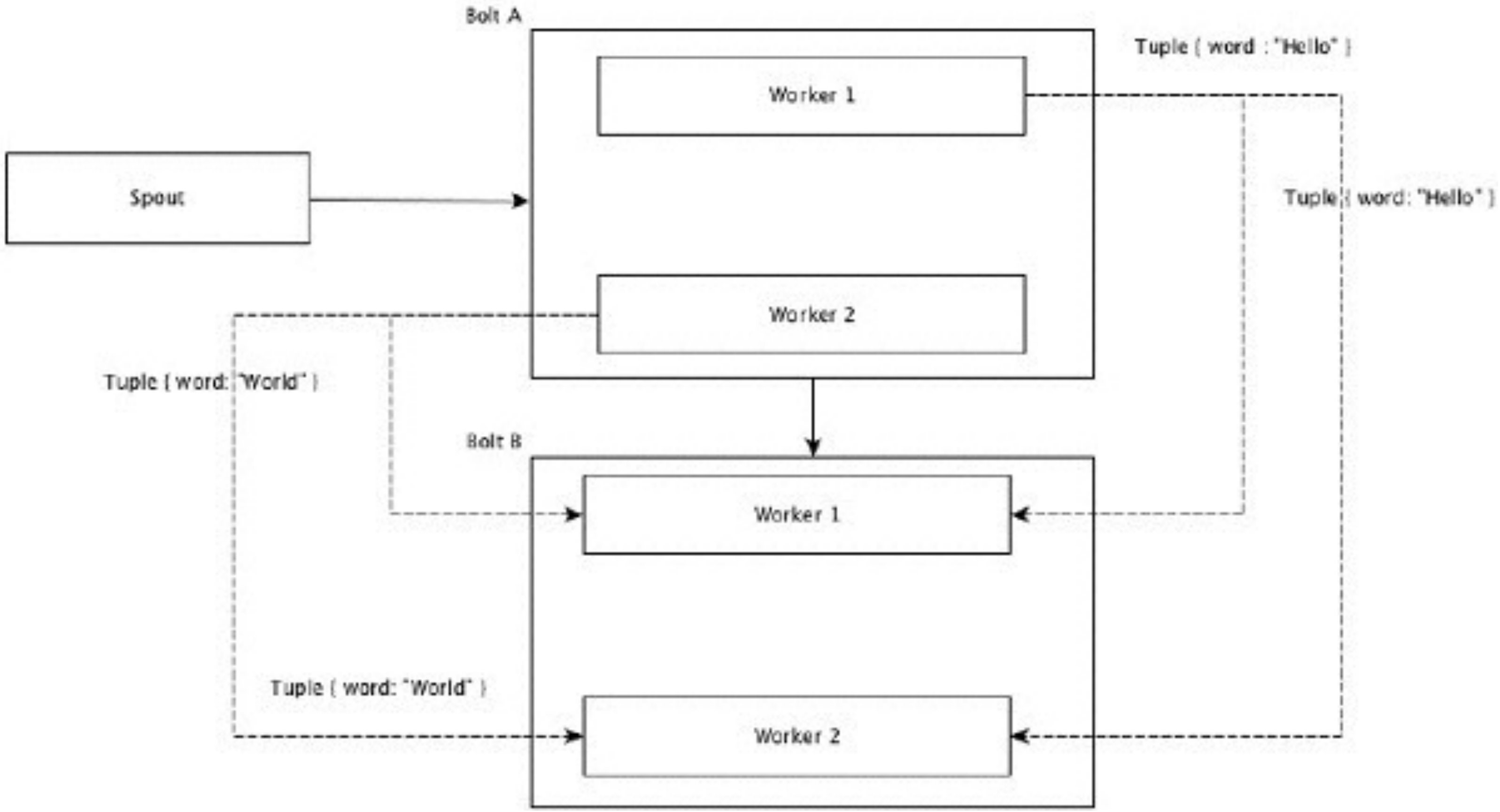
- Field Grouping- Fields with same values in tuples are grouped together and the remaining tuples kept outside.



- Global Grouping- All the streams can be grouped and forward to one bolt.



- All Grouping sends a single copy of each tuple to all instances of the receiving bolt.

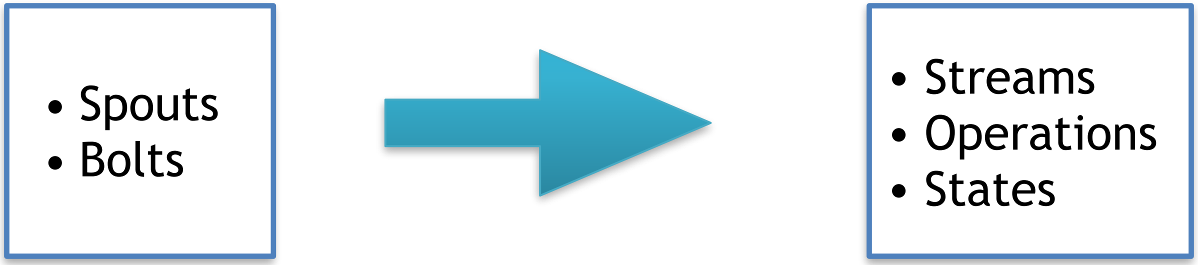


INTERVIEW TIME

- Q: Suppose I wanted to analyze words … blah blah blah … word count
- A: I shall make a topology (with which grouping?)
- Bonus: And I can also do this in real time!

Storm Trident

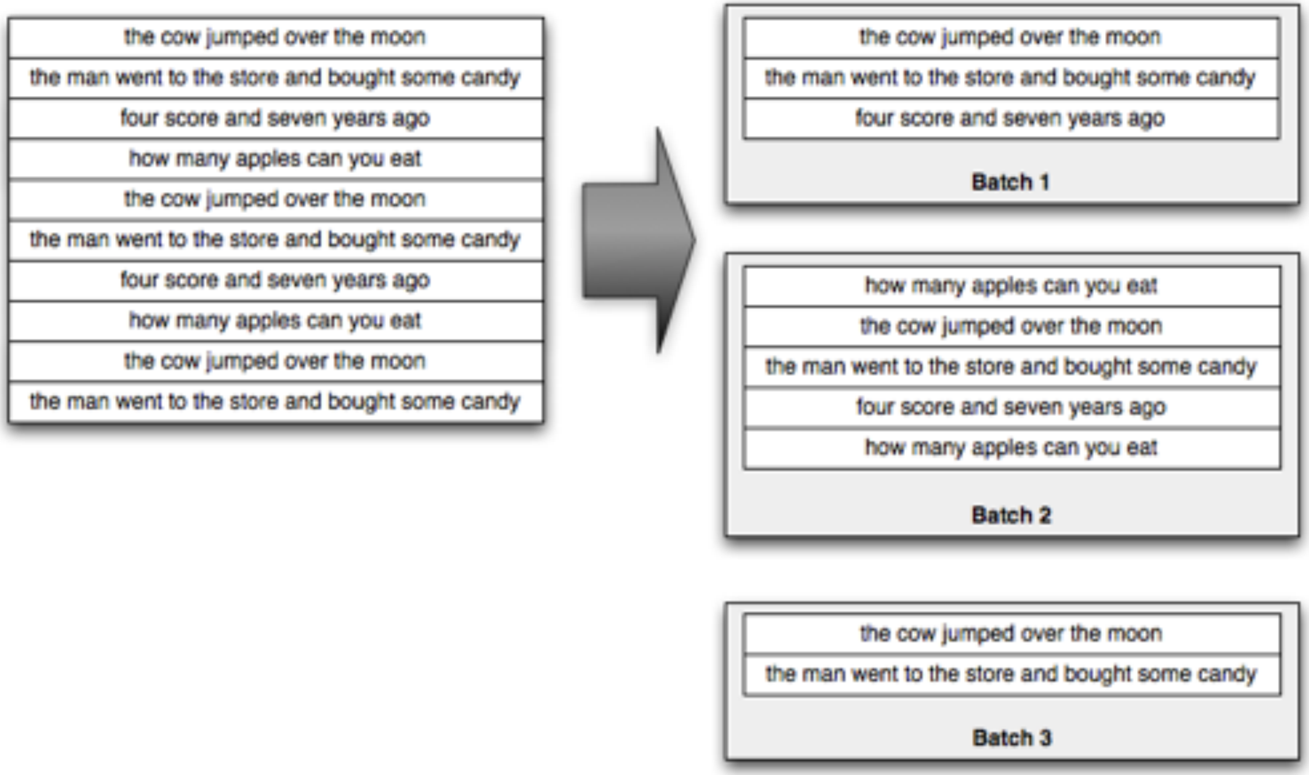
- Exactly-Once Processing
- Batch Processing
- Ordered State Updates
- Fast, Persistent Aggregation



- Spouts
- Bolts

- Streams
- Operations
- States

Batch Processing



the cow jumped over the moon
the man went to the store and bought some candy
four score and seven years ago
how many apples can you eat
the cow jumped over the moon
the man went to the store and bought some candy
four score and seven years ago
how many apples can you eat
the cow jumped over the moon
the man went to the store and bought some candy

Batch 1

the cow jumped over the moon
the man went to the store and bought some candy
four score and seven years ago

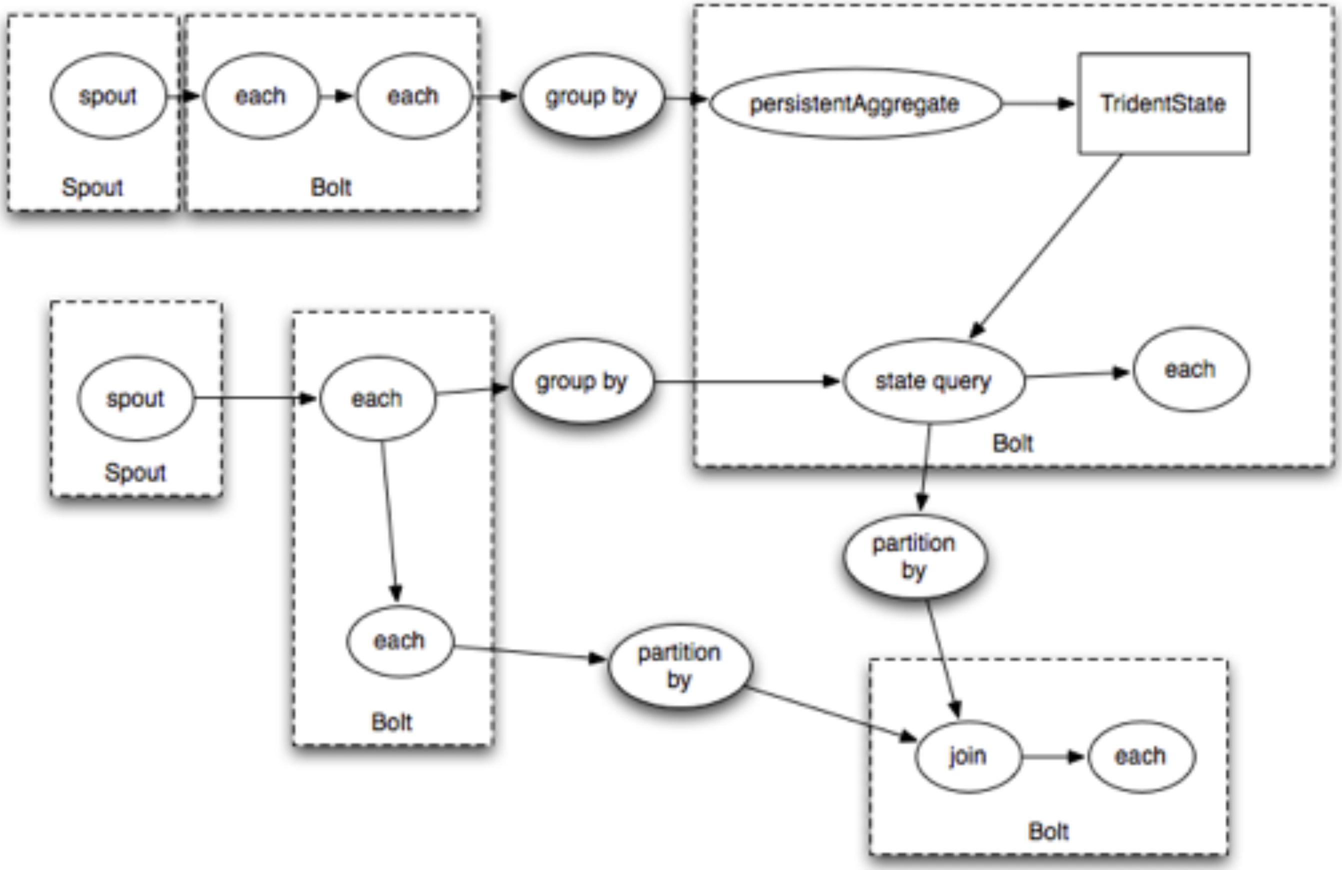
Batch 2

how many apples can you eat
the cow jumped over the moon
the man went to the store and bought some candy
four score and seven years ago
how many apples can you eat

Batch 3

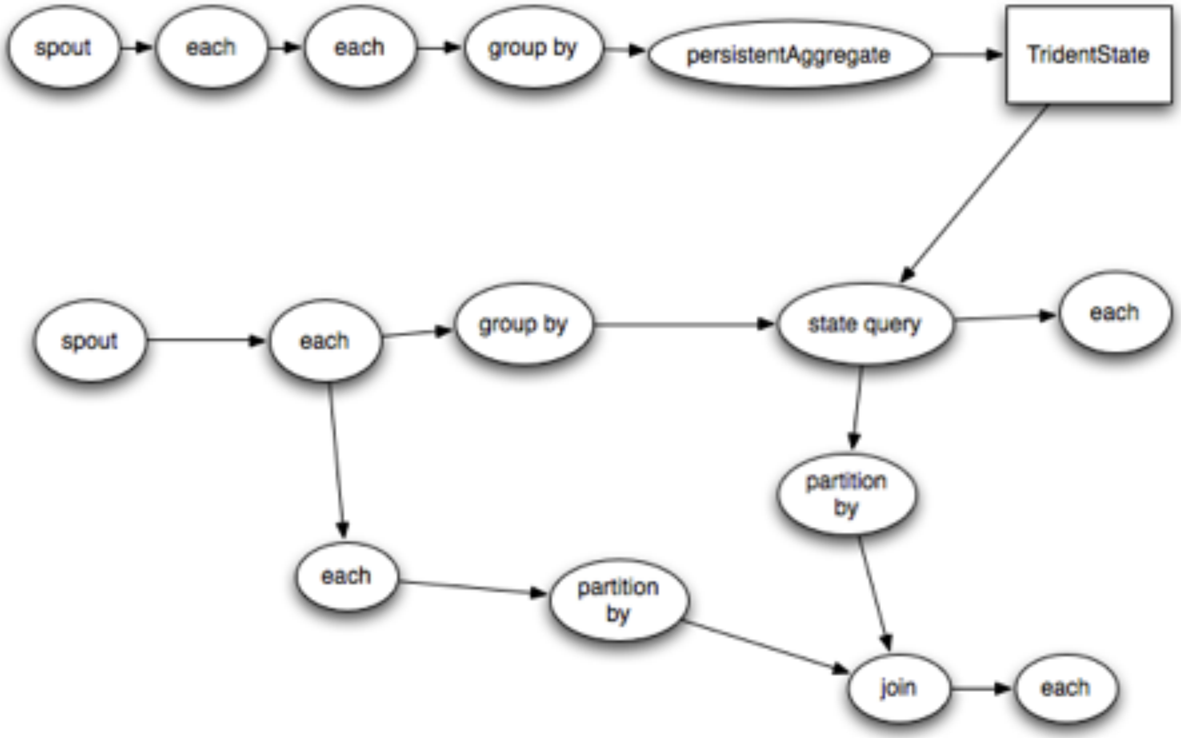
the cow jumped over the moon
the man went to the store and bought some candy

Trident Concepts



Data Application Lab

Before



Data Application Lab

Trident Demo

Trident Aggregation

- Init() Method - What is the object you are using to keep track of aggregation?
- Aggregate() Method - For each tuple, what is the aggregation process?
- Complete() Method - After the aggregation how do you output the results?

Data Application Lab

INTERVIEW TIME

- How would you process data … write into database .. integrate with this system … ETL
- A: Use Trident
- Why: Exactly Once Processing, Retries, ect.

References

- Storm Tutorial - http://storm.apache.org/releases/current/Tutorial.html
- Good Storm Overview - https://www.tutorialspoint.com/apache_storm/apache_storm_cluster_architecture.htm
- Git: https://github.com/apache/storm

Data Application Lab