Whirlpool

Data Acquisition using N-node Distributed Web Crawler

Rihan Pereira, MSCS

Advisor: Dr. Michael Soltys
Department of Computer Science
MSCS Graduate 2018-2019

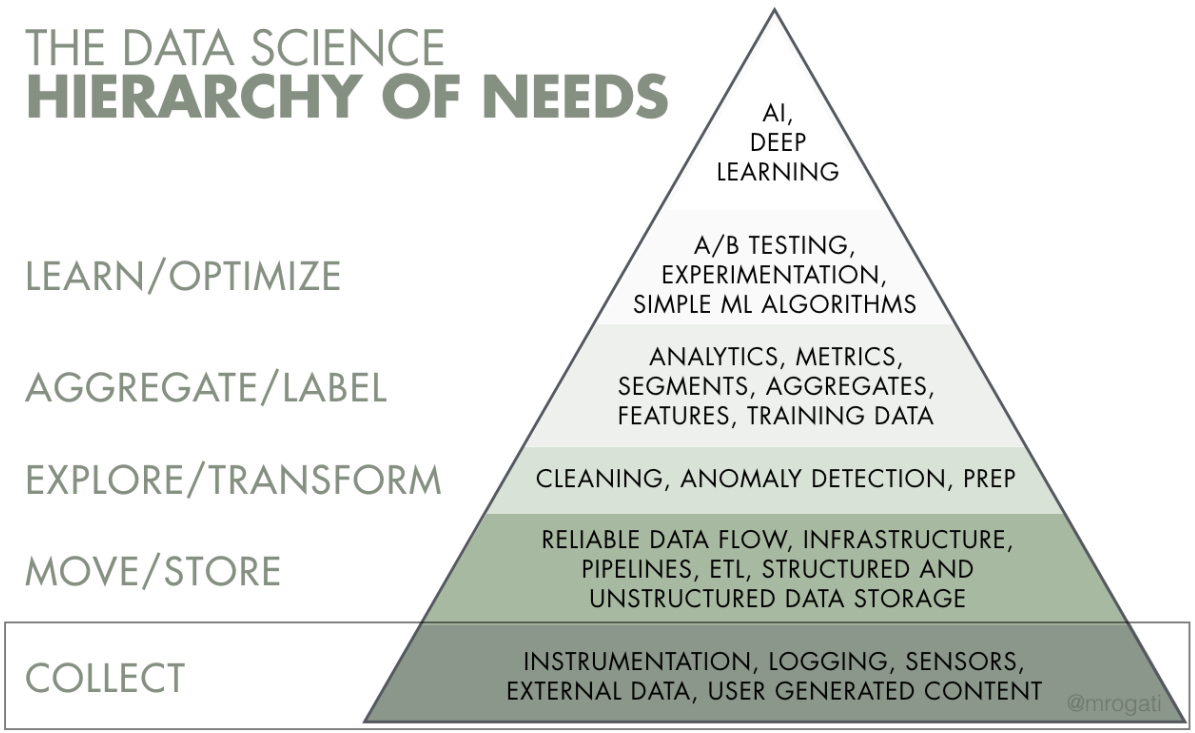
November 24, 2019

Channel Islands
CALIFORNIA STATE UNIVERSITY

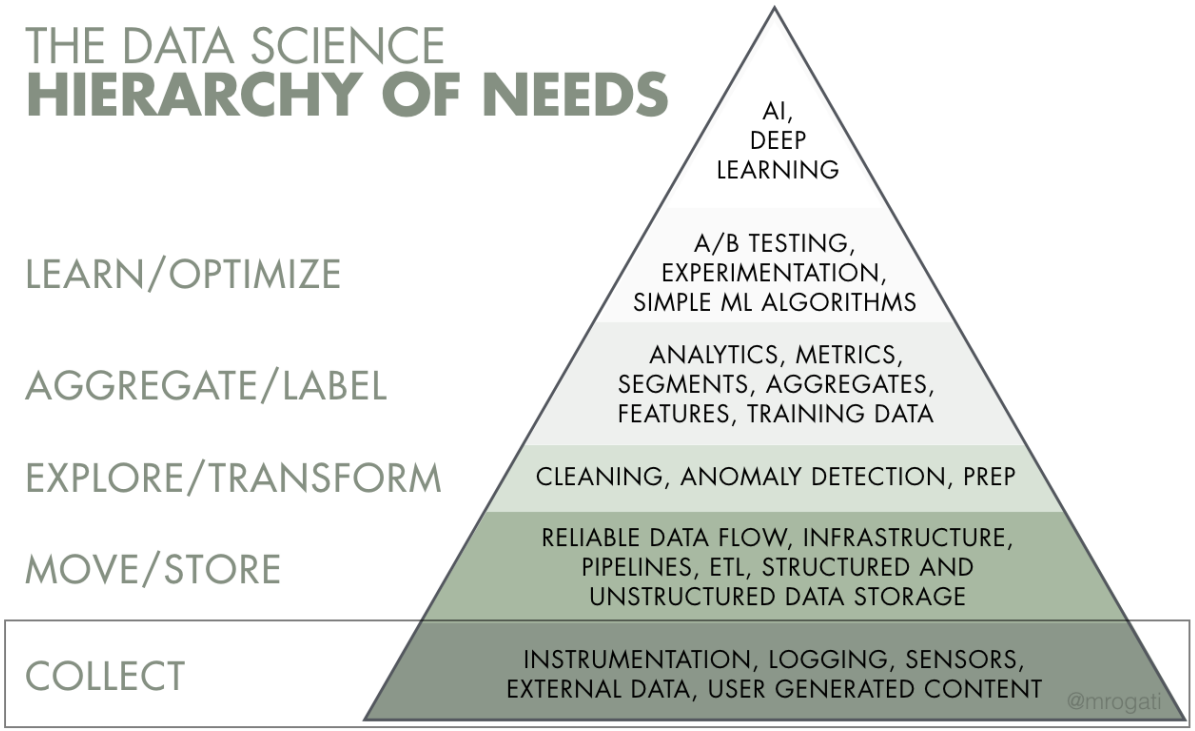
1. Motivation & Contribution
2. Crawler characteristics & history
3. Mercator 1999(Heydon & Najork)
4. Software Design Principles
5. Whirlpool: Event-driven architecture
6. Whirlpool: Parser
7. Whirlpool: Near-Deduplication
8. Whirlpool: Distributed Crawling
9. Whirlpool: Operations
10. Future work

Motivation & Contribution

Motivation


THE DATA SCIENCE
HIERARCHY OF NEEDS
AI, DEEP LEARNING
LEARN/OPTIMIZE
A/B TESTING, EXPERIMENTATION, SIMPLE ML ALGORITHMS
AGGREGATE/LABEL
ANALYTICS, METRICS, SEGMENTS, AGGREGATES, FEATURES, TRAINING DATA
EXPLORE/TRANSFORM
CLEANING, ANOMALY DETECTION, PREP
MOVE/STORE
RELIABLE DATA FLOW, INFRASTRUCTURE, PIPELINES, ETL, STRUCTURED AND UNSTRUCTURED DATA STORAGE
COLLECT
INSTRUMENTATION, LOGGING, SENSORS, EXTERNAL DATA, USER GENERATED CONTENT
@mrogati

Motivation



Self-actualization (AI) is great, but you first need food, water, and shelter (data literacy, collection, and infrastructure).”

Contributions

to be completed

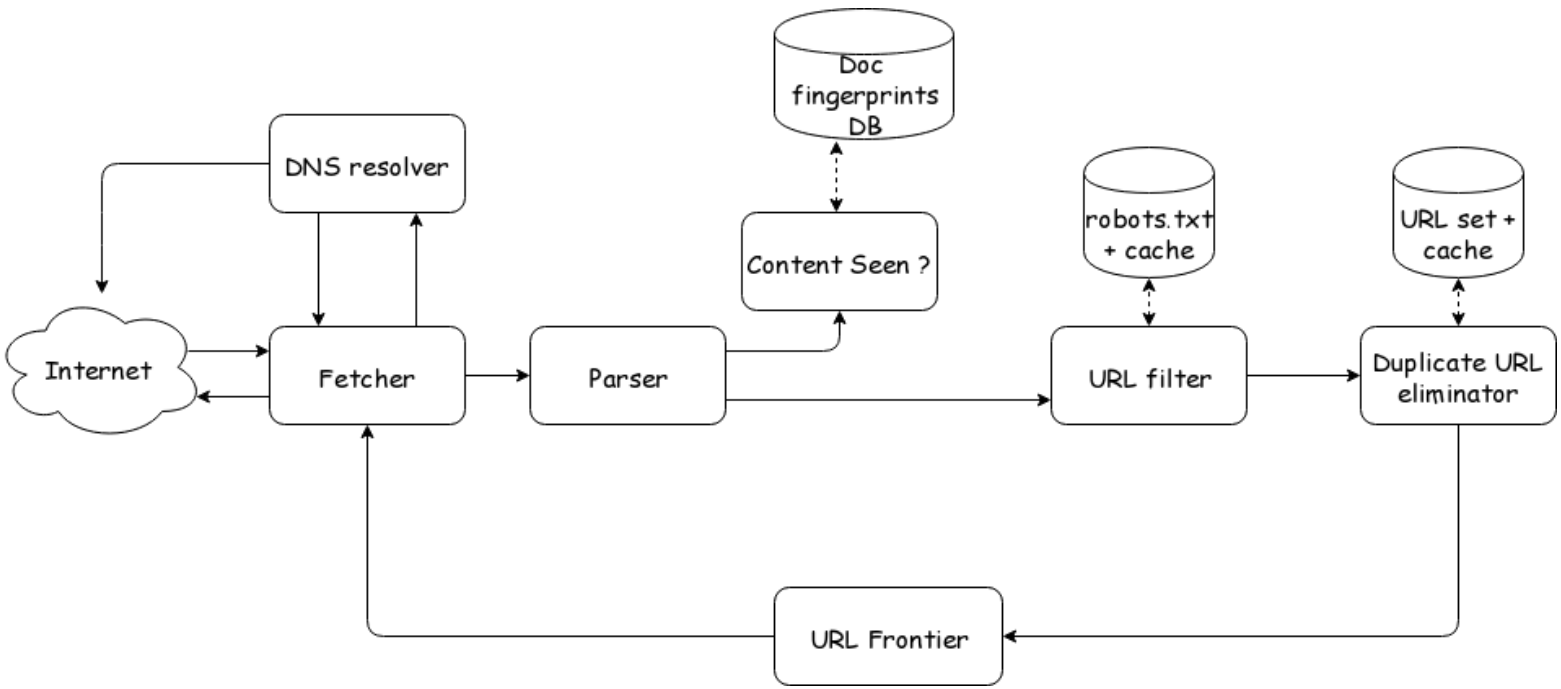
Crawler characteristics & history

Crawl char & hist.

to add something

Mercator 1999(Heydon & Najork)

Mercator background



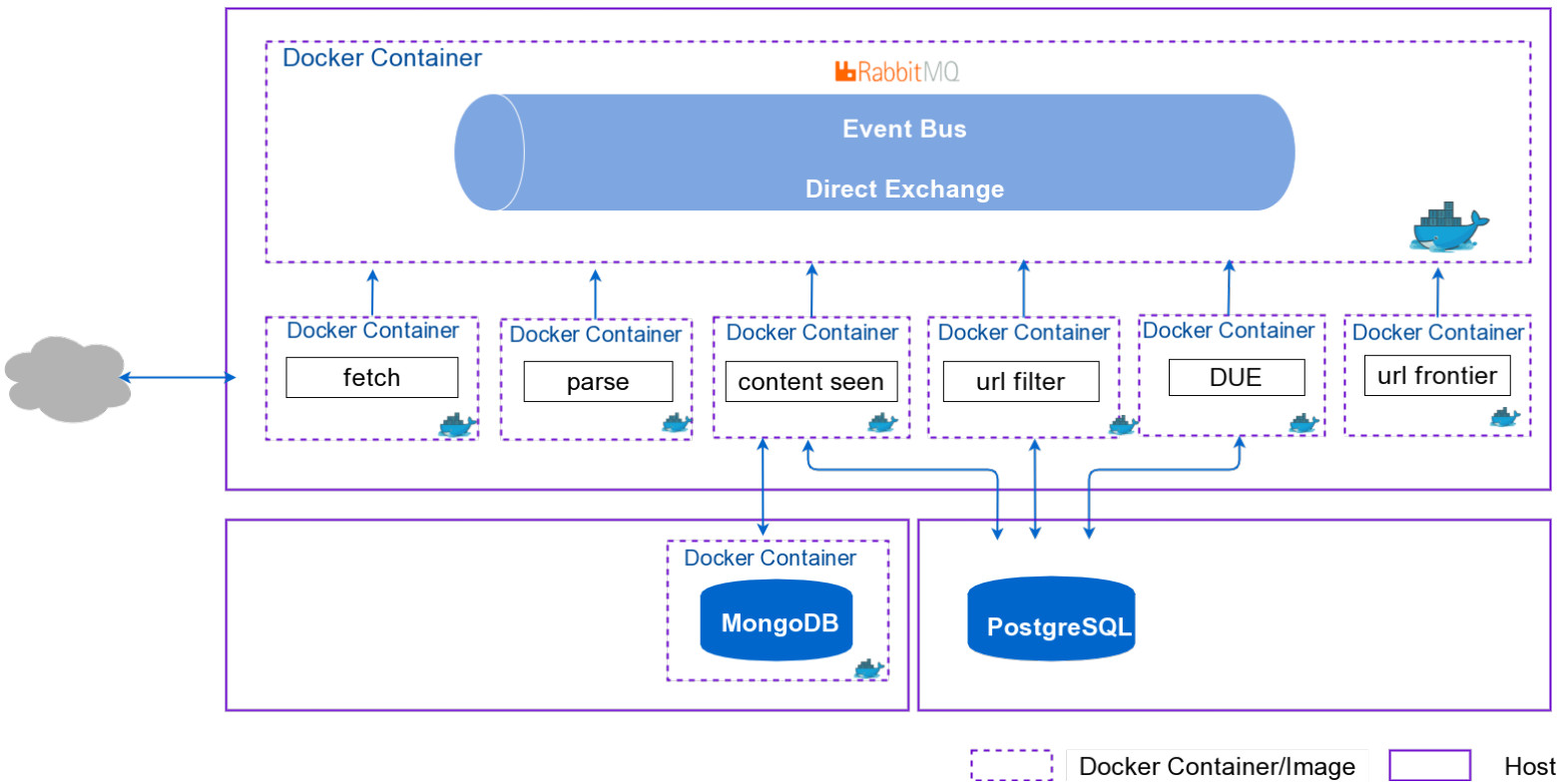
Software Design Principles

Soft. Design

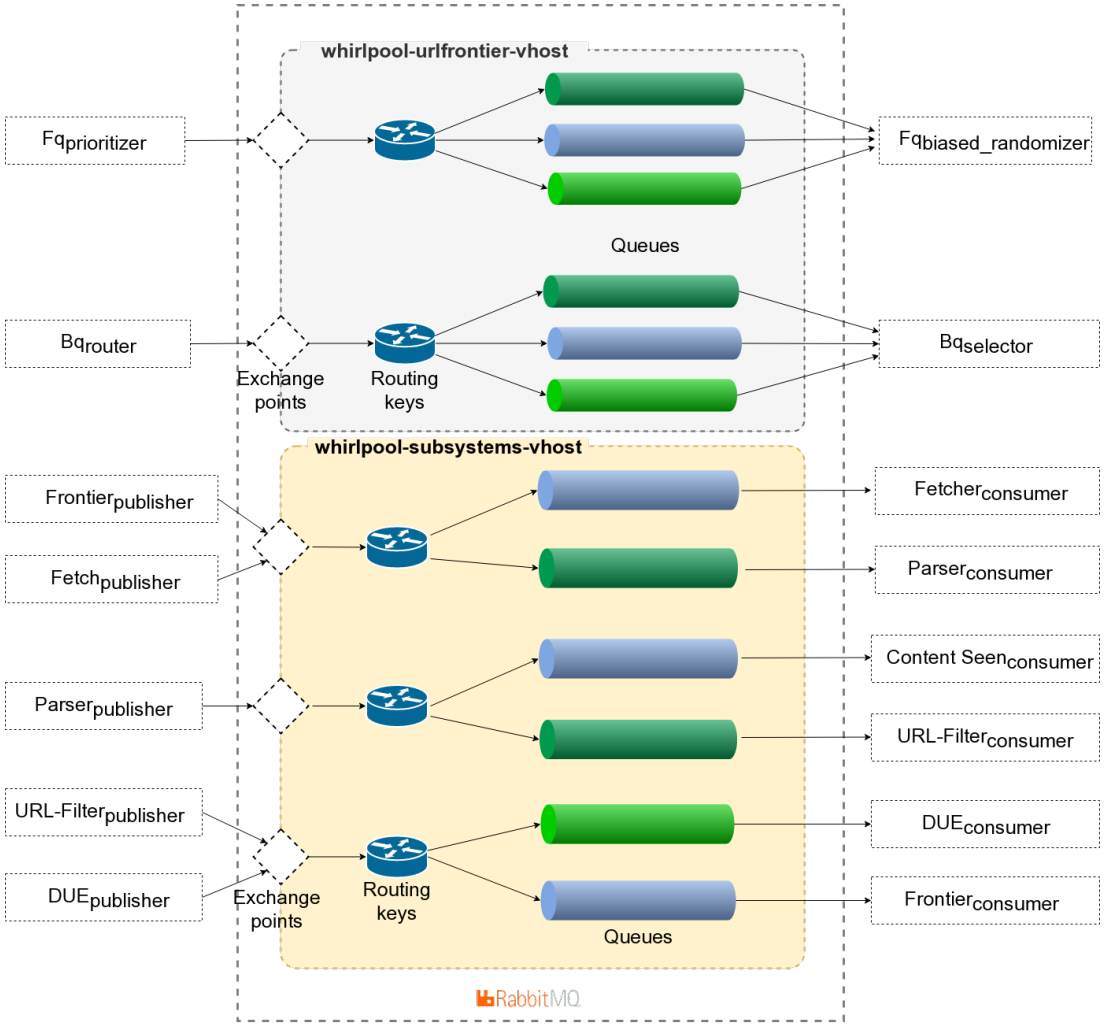
to add something

Whirlpool: Event-driven architecture

RabbitMQ: Message bus



Direct Worker Queue Data Flow



Whirlpool: Parser

Parser

to add something

Whirlpool: Near-Deduplication

Dedupe

to add something

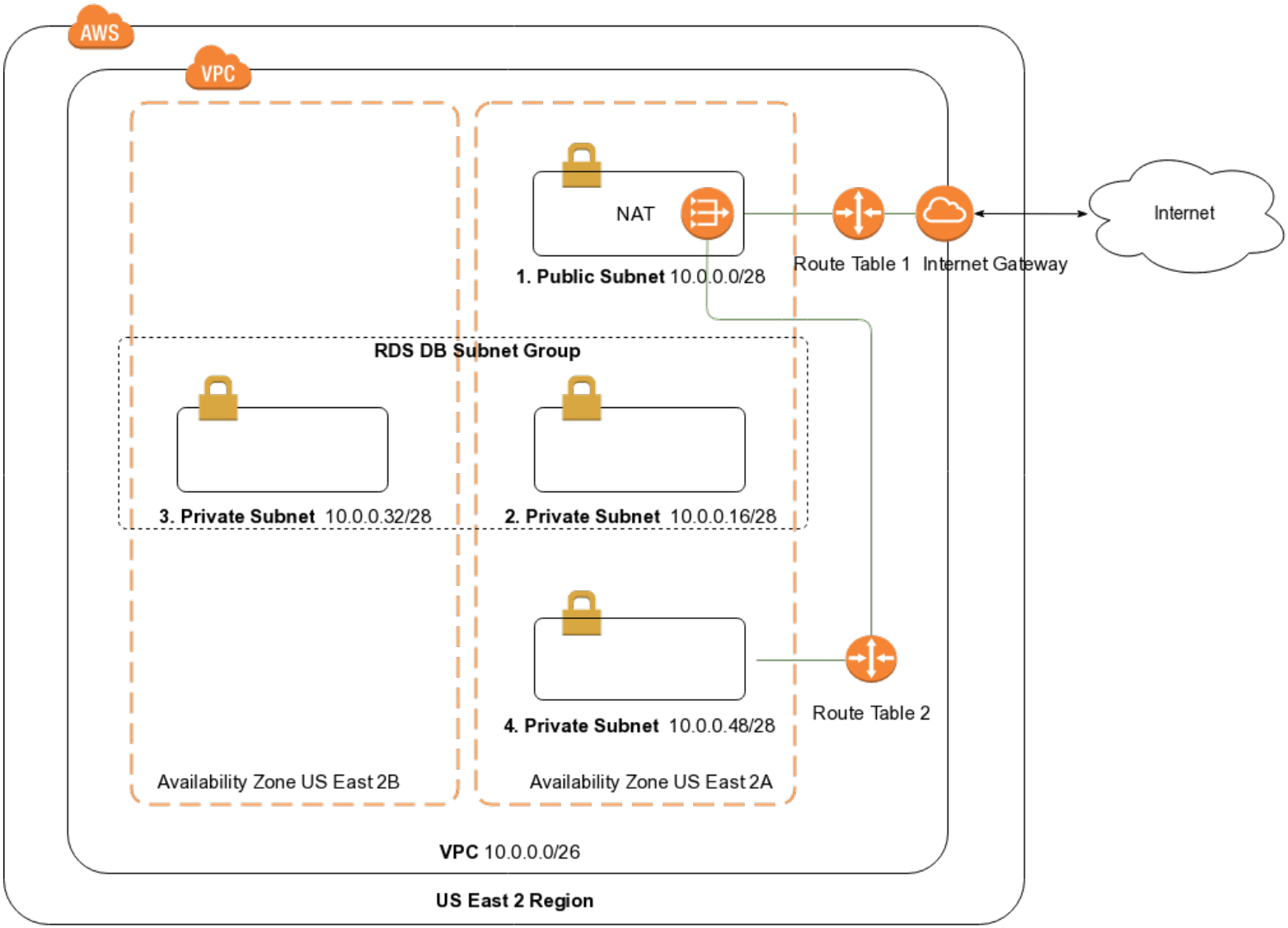
Whirlpool: Distributed Crawling

Dist. crawl

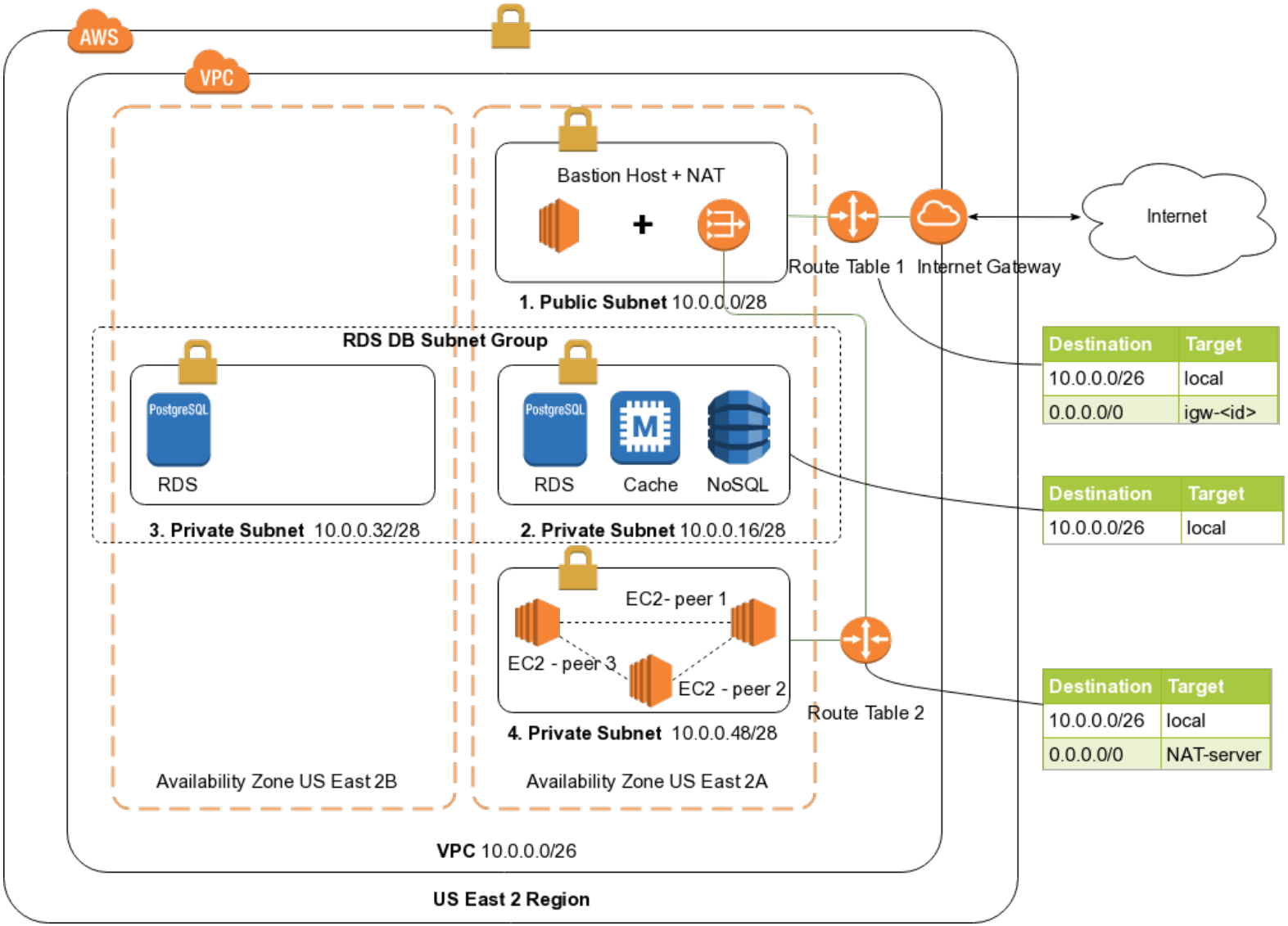
to add something

Whirlpool: Operations

From 10,000 ft.



From 5,000 ft.


AWS
VPC
Bastion Host + NAT
1. Public Subnet 10.0.0.0/28
Route Table 1
Internet Gateway
Internet
RDS DB Subnet Group
PostgreSQL
RDS
3. Private Subnet 10.0.0.32/28
PostgreSQL
RDS
Cache
NoSQL
2. Private Subnet 10.0.0.16/28
EC2- peer 1
EC2 - peer 3
EC2 - peer 2
4. Private Subnet 10.0.0.48/28
Route Table 2
Availability Zone US East 2B
Availability Zone US East 2A
VPC 10.0.0.0/26
US East 2 Region
Destination Target
10.0.0.0/26 local
0.0.0.0/0 igw-<id>
Destination Target
10.0.0.0/26 local
Destination Target
10.0.0.0/26 local
0.0.0.0/0 NAT-server

Future work

future to do

to add something

Thank you! Questions ?