BATCH

(TO)

NEAR-REALTIME

Shivji kumar Jha, Staff Engineer, Nutanix
Sachidananda Maharana, MTS 4, Nutanix

Legal Disclaimer

Forward Looking Statements

This presentation and the accompanying oral commentary may contain express and implied forward-looking statements, including, but not limited to, statements relating to: our business plans, initiatives and objectives; ability to execute such plans, initiatives and objectives in a timely manner, and the benefits and impact of such plans, initiatives and objectives, including our ability to manage our expenses in future periods; our financial targets and our plans to achieve those targets; the benefits and capabilities of our platform, products, services and technology; our plans and expectations regarding new products, services, product features and techno including those that are still under development or in process; the timing of any product releases or upgrades or announcements; anticipated trends, growth rates and challenges in our business and in the markets in whi operate; our ability to develop new solutions, product features and technology and bring them to market in a timely manner, as well as the impact and/or benefits of including additional solutions or features in our produ portfolio; market acceptance of new technology and recently introduced solutions; the interoperability and availability of our solutions with and on third-party platforms, including public cloud platforms; our ability to ma and strengthen our relationships with our channel partners, OEMs and other third parties, and the impact of any changes to such relationships on our business, operations and financial results; the competitive market, inc our competitive position and ability to compete effectively, our projections about our market share in future periods, the competitiveness of our future cost structure with those of other companies, and the competitive advantages of our products; our plans and timing for, and the success and impact of, our transition to a subscription-based business model; and macroeconomic trends and geopolitical environment, including the on-go global supply chain disruptions.

These forward-looking statements are not historical facts and instead are based on our current expectations, estimates, opinions, and beliefs. Forward-looking statements should not be considered as guarantees or pred of future events. Consequently, you should not rely on these forward-looking statements. The accuracy of these forward-looking statements depends upon future events and involves risks, uncertainties, and other factor including factors that may be beyond our control, that may cause these statements to be inaccurate and cause our actual results, performance or achievements to differ materially and adversely from those anticipated o implied by such statements, including, among others, the risks detailed in our most recent Annual Report on Form 10-K and Quarterly Reports on Form 10-Q, each as filed with the U.S. Securities and Exchange Commissi the SEC, which should be read in conjunction with the information in this presentation and accompanying oral commentary. Our SEC filings are available on the Investor Relations section of our website at ir.nutanix.com a the SEC's website at www.sec.gov. These forward-looking statements speak only as of the date of this presentation and accompanying oral commentary and, except as required by law, we assume no obligation, and exp disclaim any obligation, to update, alter or otherwise revise any of these forward-looking statements to reflect actual results or subsequent events or circumstances.

Product or Roadmap Information

Any future product or roadmap information included in this presentation and the accompanying oral commentary is (i) intended to outline general product directions, (ii) not a commitment, promise or legal obligation fo Nutanix to deliver any material, code, or functionality, and (iii) not intended to be, and shall not be deemed to be, incorporated into any contract. This information should not be used when making a purchasing decision. note that Nutanix has made no determination as to whether separate fees will be charged for any future products, product enhancements and/or functionality which may ultimately be made available. Nutanix may, in its discretion, choose to charge separate fees for the delivery of any future products, product enhancements and/or functionality which are ultimately made available.

Third Party Reports and Publications

Certain information contained in the presentation and accompanying oral commentary made available as part of this digital event may relate to or be based on reports, studies, publications, surveys and other data obtai from third-party sources and our own internal estimates and research. While we believe these third-party reports, studies, publications, surveys and other data are reliable as of the date of the applicable presentation, the not independently verified, and we make no representation as to the adequacy, fairness, accuracy, or completeness of any information obtained from third-party sources.

Trademark Disclaimer

ABOUT US

Shivji Kumar Jha
Staff Engineer
CPaaS Data Platform,
Nutanix

- Software Engineer & Regular Speaker / Meetups
- Excited about:
  - Distributed Databases & Streaming
  - Open-Source Software & Communities
  - MySQL/Postgres, Pulsar/NATS, Druid/Clickhouse

Sachidananda Maharana
Software Engineer
OLAP Ninja
CPaaS Team, Nutanix

- Regular Platform Engineer
- Excited about:
  - Distributed OLAP Databases
  - Open-Source Enthusiast

CONTENTS

- Background
  - App Context
  - The problem Statement
  - Druid 101
- Ingestion Issues & Fixes
- Query Issues & Fixes
- Fix in common libraries (OSS)
  - Be a good citizen!

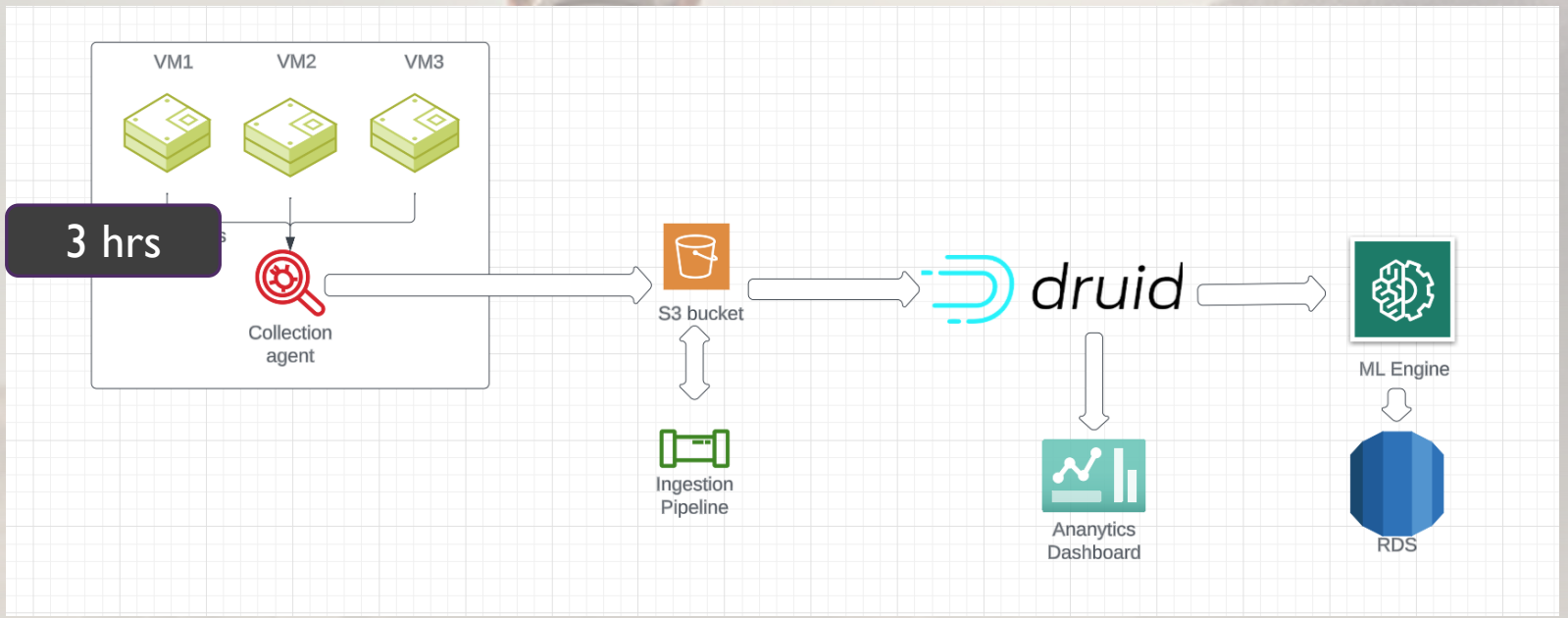
BACKGROUND: APPLICATION CONTEXT

- UI based slice and dice analytics with filters
- Easy Ops: One Multi-tenant DB (Druid) Cluster
- Fine grained isolation per customer / use-case
- Resilient Pipeline: Temporal for orchestration
- Durable / Scalable Storage: S3
- Java Workers with Postgres storage for state.

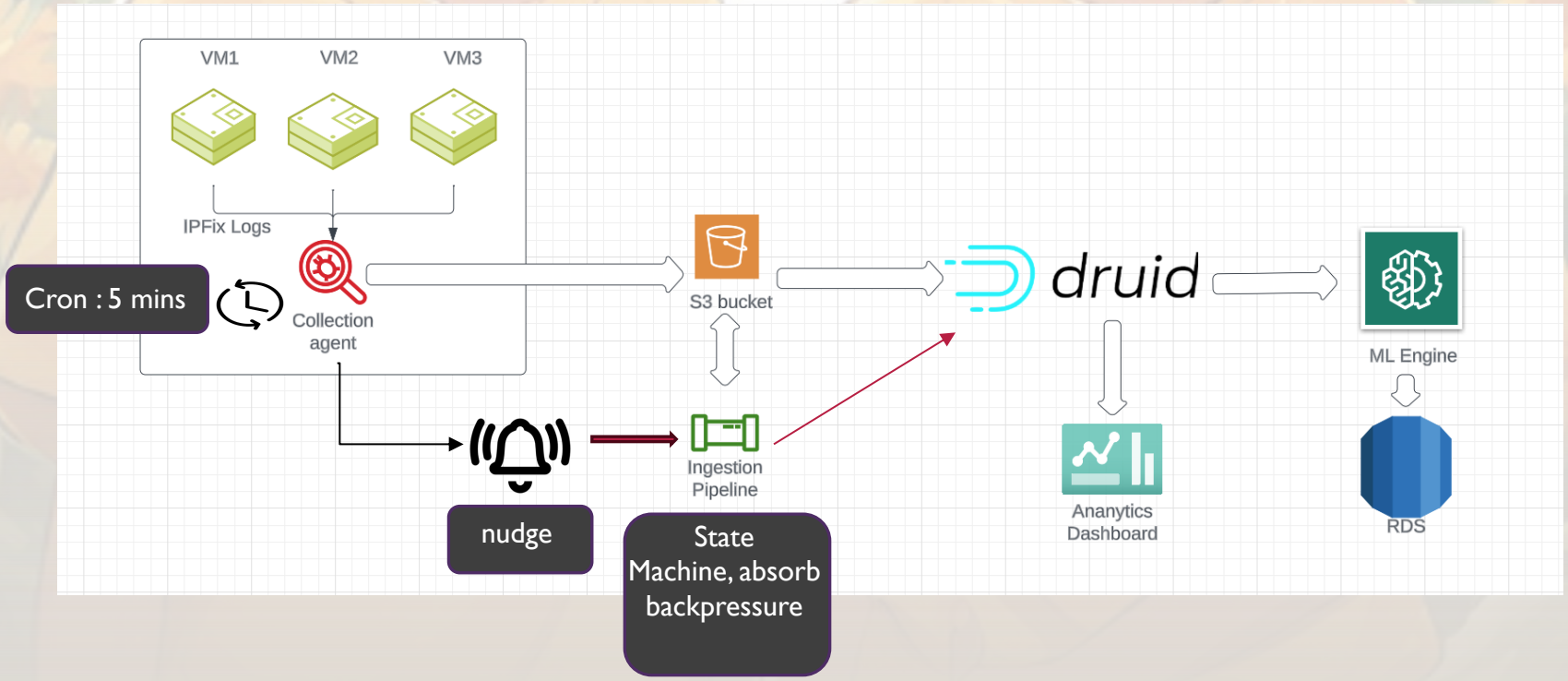
BACKGROUND: APP INGESTION PIPELINE

- IPFIX log files are collected from clouds.
  - IPFIX : IP Flow Information Export
  - Summarizes network data packets to track IP actions
- We enrich data and store in an s3 bucket.
- S3 data is ingested into druid.
- Serves Analytics dashboards in slice and dice manner.
- Used for ML engine as well.

EARLIER: OLD BATCH SYSTEM



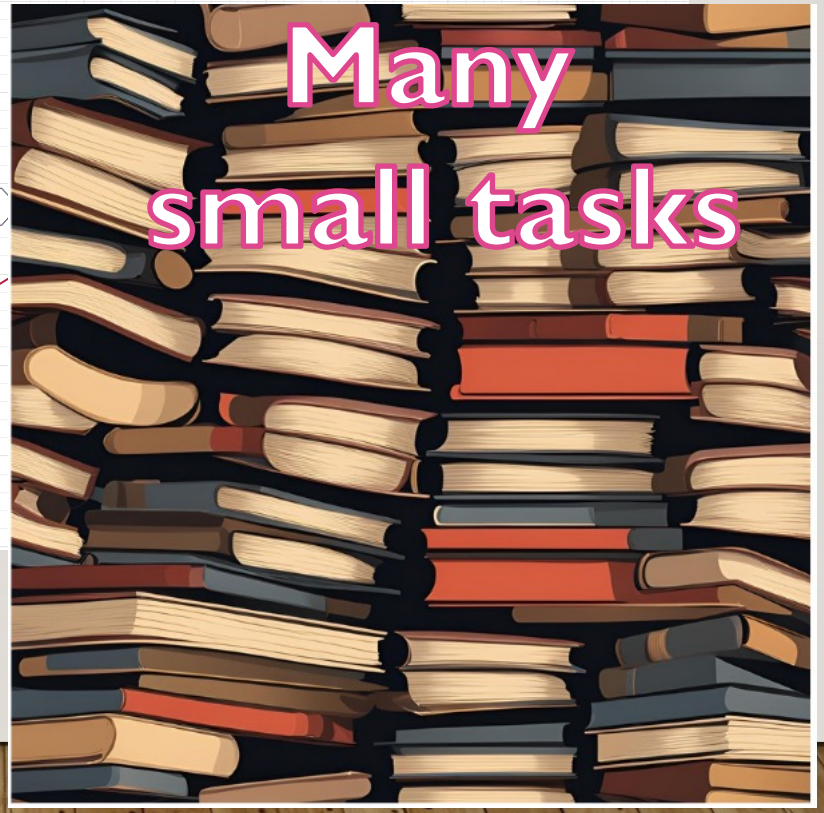
NOW: NEAR-REAL-TIME INGESTION



BATCH TO NEAR-REAL-TIME SYSTEM

Druid Ingestion Tasks

Ingestion V1



Ingestion V2

ANY IDEAS HOW TO HANDLE THIS?

https://www.dreamstime.com/stock-image-diversity-show-hands-image9264971

IDEAS?

- Scale [UP / OUT] resources?
- Isolate tasks [ingest / compact / query]?
- Parallelize?
- Choose simple performant data assignment algos?
- Throttle & throw defined error codes
  - Global Queues
  - Per customer

MORE IDEAS: COMPACTION FOR BETTER QUERY!

- A lot of task means a lot of small files
- During query, reading from lots of small files not efficient
- S3 also prefers smaller number of bigger files.
- Solution: Merge files in the background
- Enter Compaction:
  - Combine smaller files into rolled up (fewer) bigger files.
  - Reduce the granularity as data gets old!
    - Example: Minute => hourly => daily => weekly…
    - Can't graphs for huge time at less granularity.
    - Saves disk space, network throughput, latency

SUMMARY: CHANGE IN REQUIREMENTS

- Change in Requirement: Batch (3 hours) to 5 minutes
- Earlier:
  - Agent collects data, dumps to S3.
  - Cron runs every 3 hour, ingests from S3 to Druid
  - SLA : 3 hours
- New Design:
  - SLA : 10 minutes
  - Agent collects data, dumps to S3 every 5 minutes.
  - Ingestion Pipeline ingests to Druid depending on what Druid likes.
  - Ingestion Pipeline gobbles backpressure.
- Release Plan
  - Data sources uploaded to cluster in a phased manner

HEARD OF DRUID?

*https://www.dreamstime.com/stock-image-diversity-show-hands-image9264971

DRUID 101

Open-source, Apache 2.0 License and under Apache Foundation

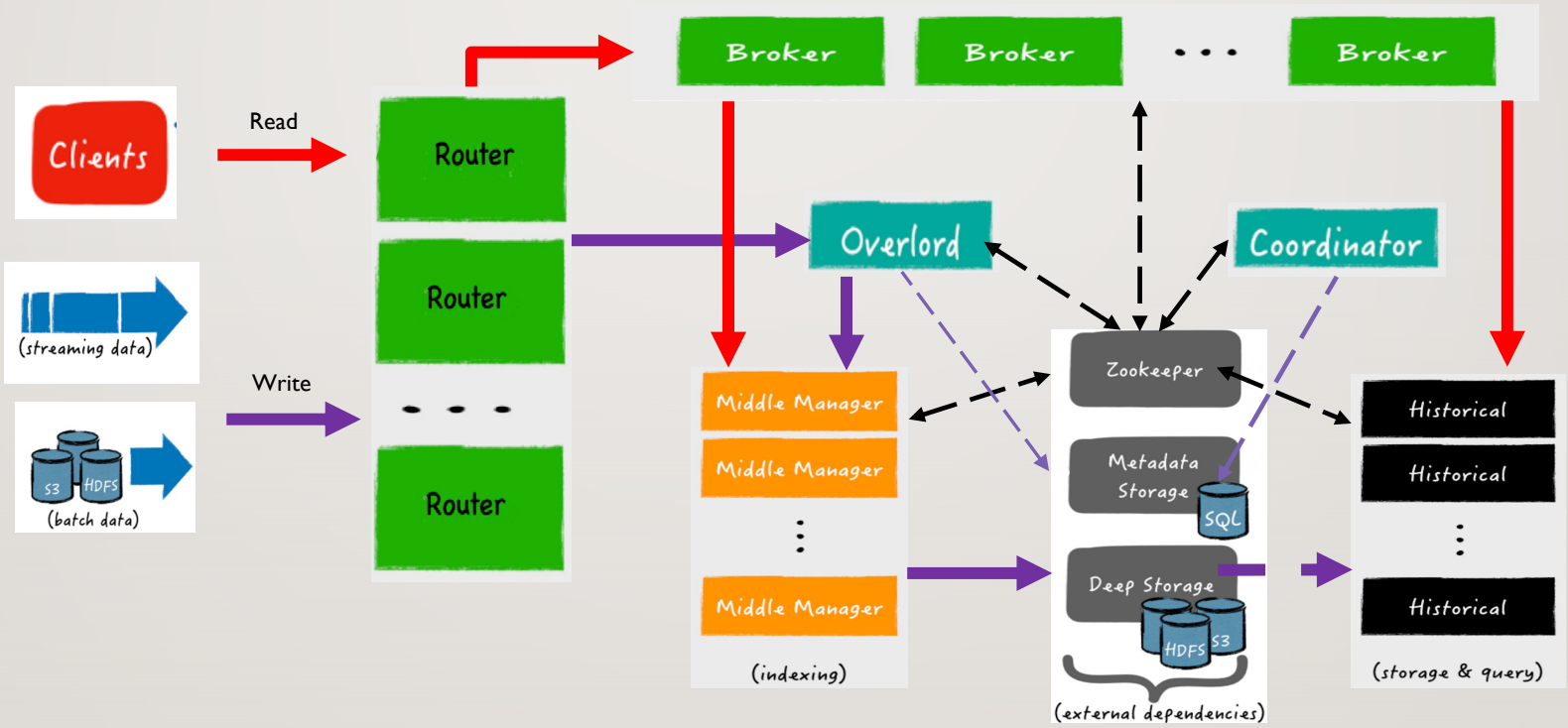
Columnar data store designed for high-performance.

Supports Real-time and Batch ingestion.

Segment Oriented Storage

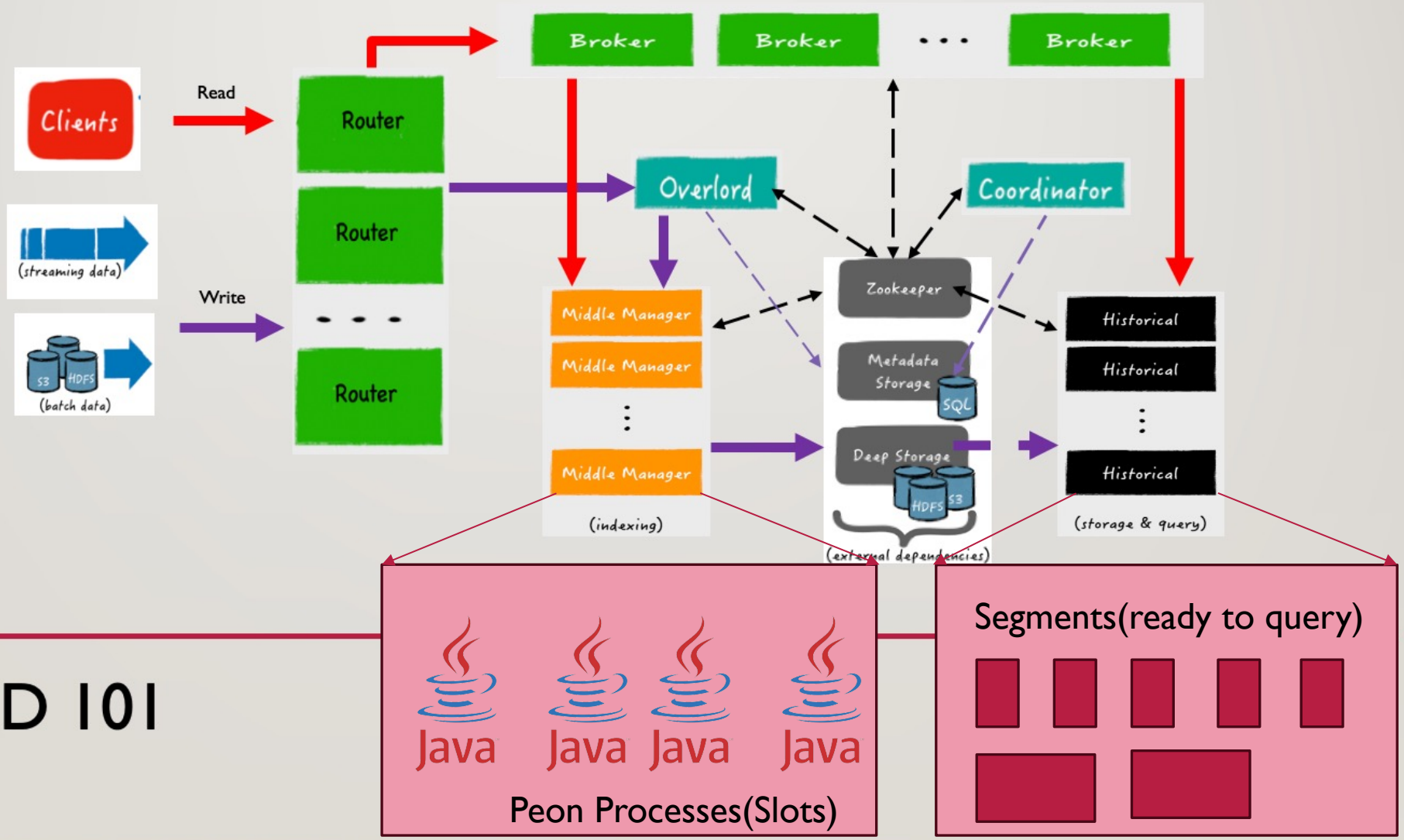
Distributed and modular architecture, horizontally scalable for most parts

Supports Data tiering - Keep cold data in cheaper storage!


Clients
Read
(streaming data)
Write
S3
HDFS
(batch data)
Router
Router
Router
Broker
Broker
Broker
Overlord
Coordinator
Middle Manager
Middle Manager
Middle Manager
(indexing)
Zookeeper
Metadata Storage
SQL
Deep Storage
HDFS
S3
(external dependencies)
Historical
Historical
Historical
(storage & query)

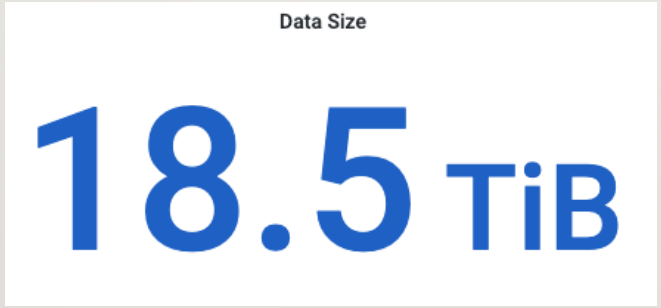
DRUID 101

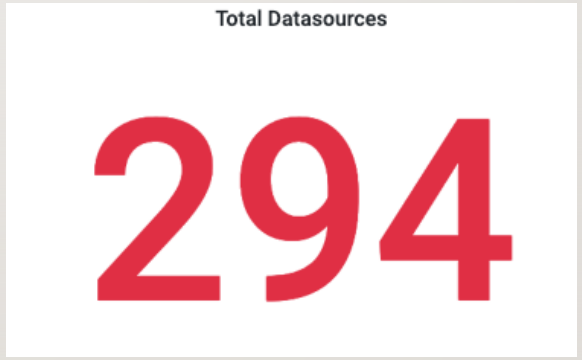
DRUID 101


Clients
Read
(streaming data)
Write
S3
HDFS
(batch data)
Router
Router
Router
Broker
Broker
Broker
Overlord
Coordinator
Zookeeper
Middle Manager
Middle Manager
Middle Manager
(indexing)
Metadata Storage
SQL
Deep Storage
HDFS
S3
(external dependencies)
Historical
Historical
Historical
(storage & query)
Java
Java
Java
Java
Peon Processes(Slots)
Segments(ready to query)

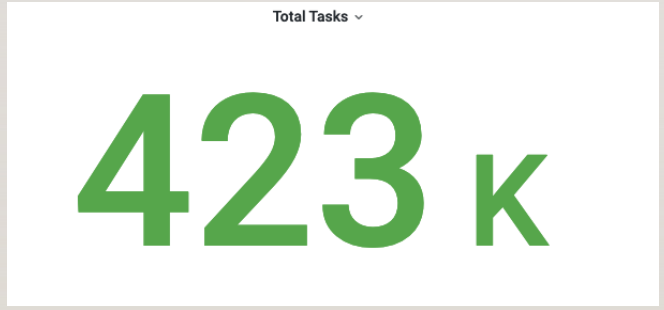
DRUID NOS : 4+ YEARS IN PROD

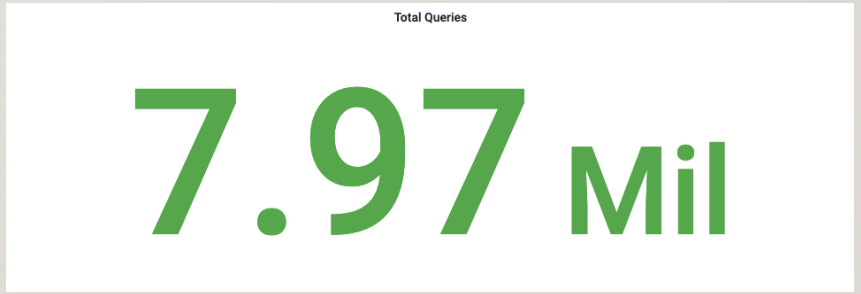
Cluster size

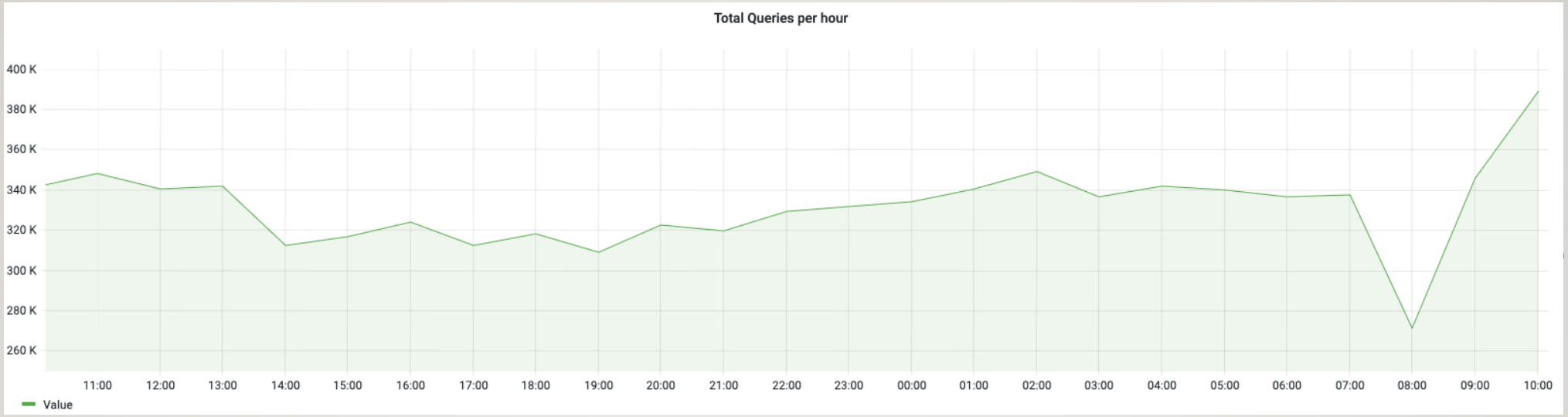


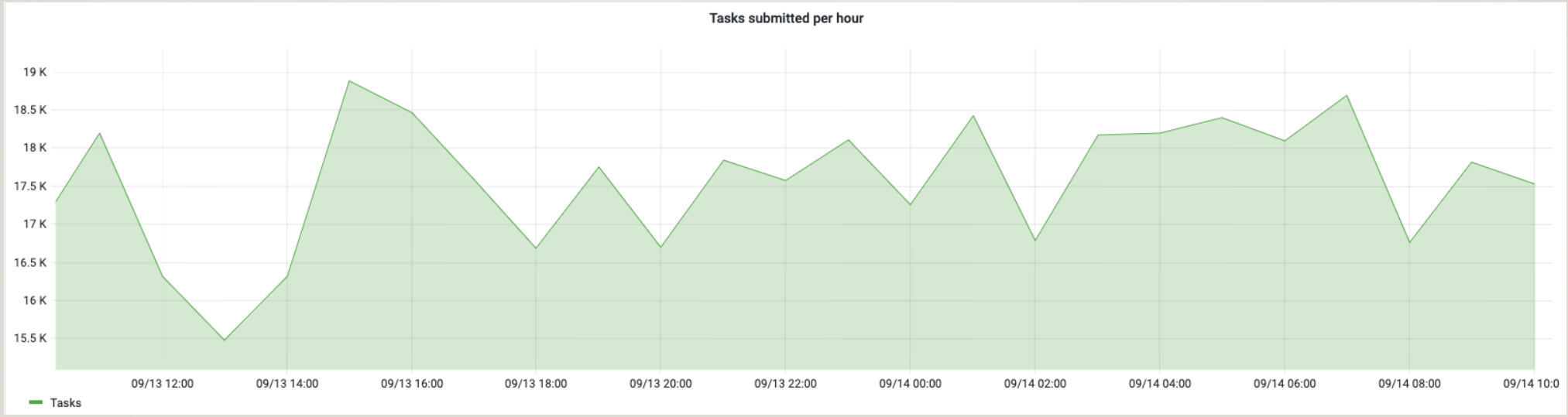


Last 24 hrs



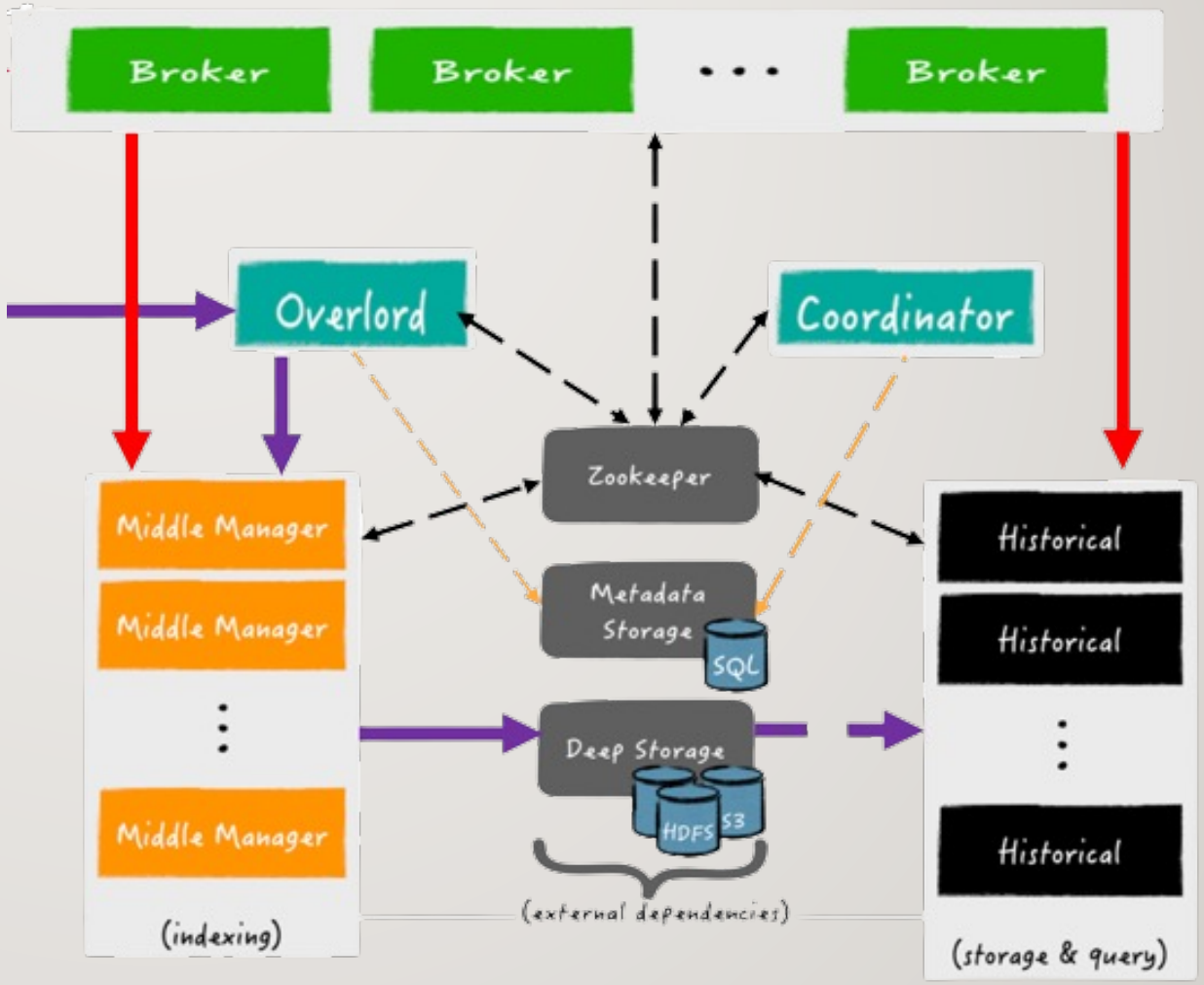



Total Queries per hour
400 K
380 K
360 K
340 K
320 K
300 K
280 K
260 K
11:00
12:00
13:00
14:00
15:00
16:00
17:00
18:00
19:00
20:00
21:00
22:00
23:00
00:00
01:00
02:00
03:00
04:00
05:00
06:00
07:00
08:00
09:00
10:00
Value


Tasks submitted per hour
19 K
18.5 K
18 K
17.5 K
17 K
16.5 K
16 K
15.5 K
09/13 12:00
09/13 14:00
09/13 16:00
09/13 18:00
09/13 20:00
09/13 22:00
09/14 00:00
09/14 02:00
09/14 04:00
09/14 06:00
09/14 08:00
09/14 10:0
Tasks

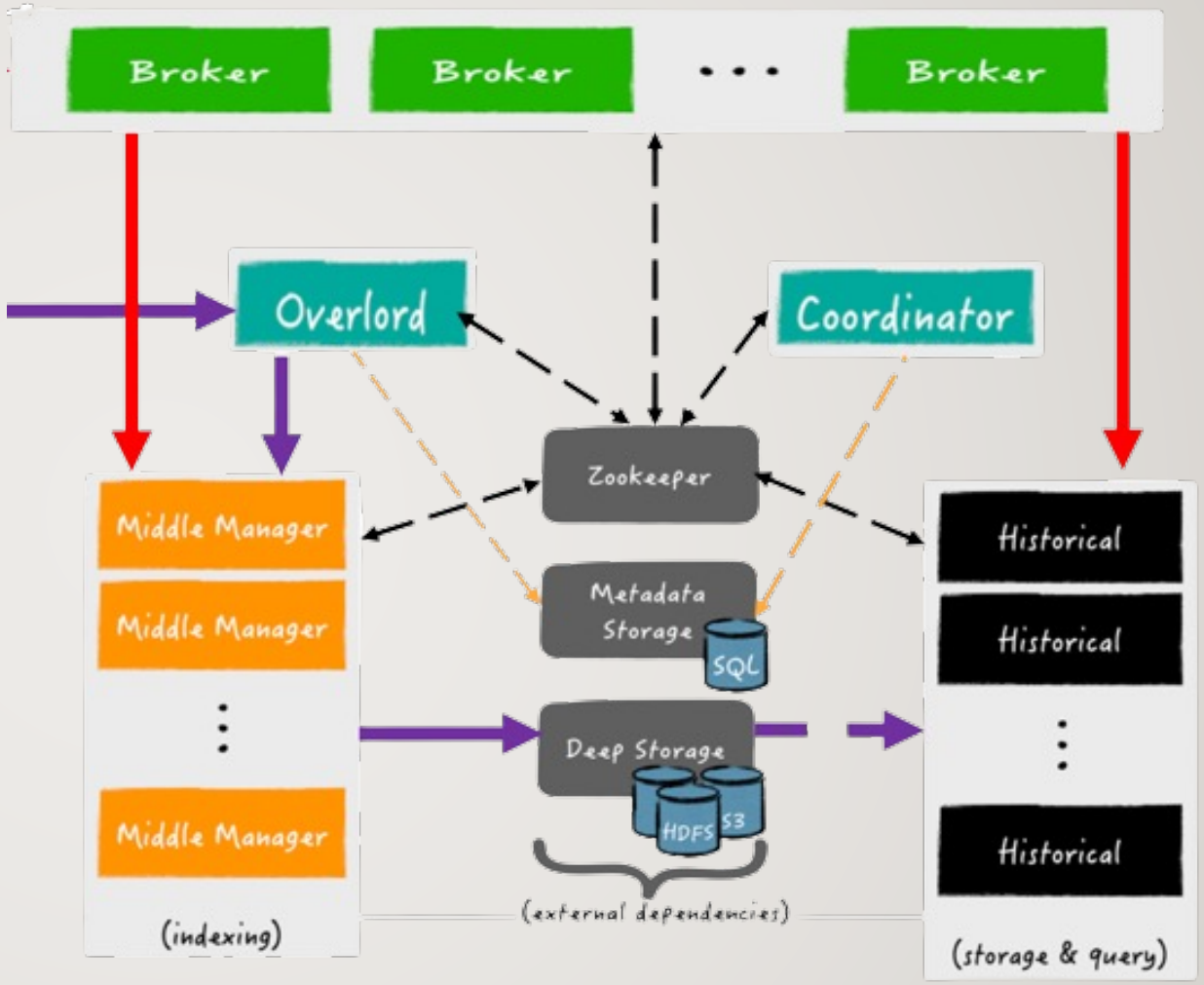
INGESTION ISSUES

DEPLOYING NEW DESIGN



DEPLOY FOR 1 CUSTOMER

Customer 1

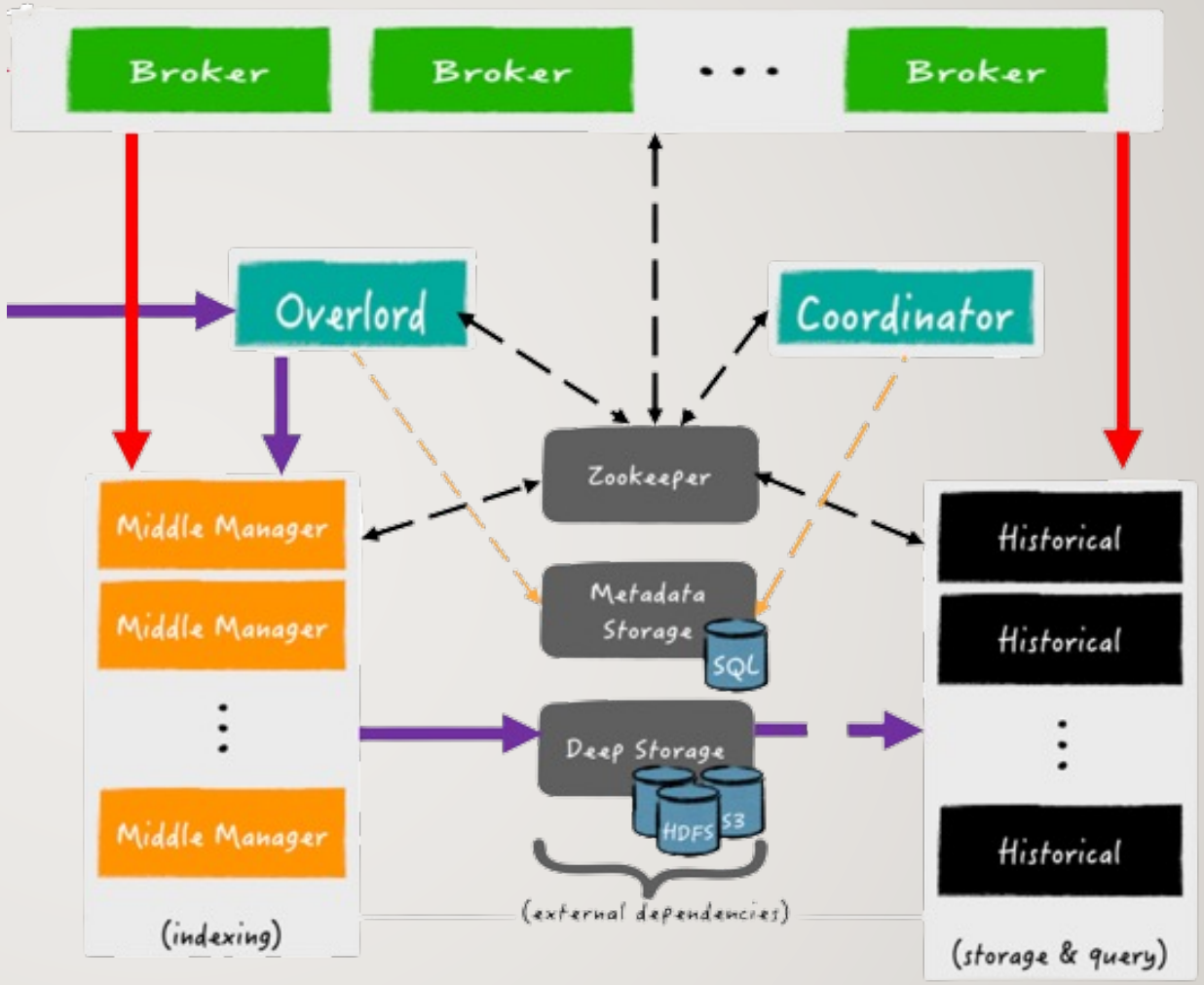


DEPLOY FOR MORE

Customer 3

Customer 2

Customer 1



DEPLOY FOR ALL…BOOOM!

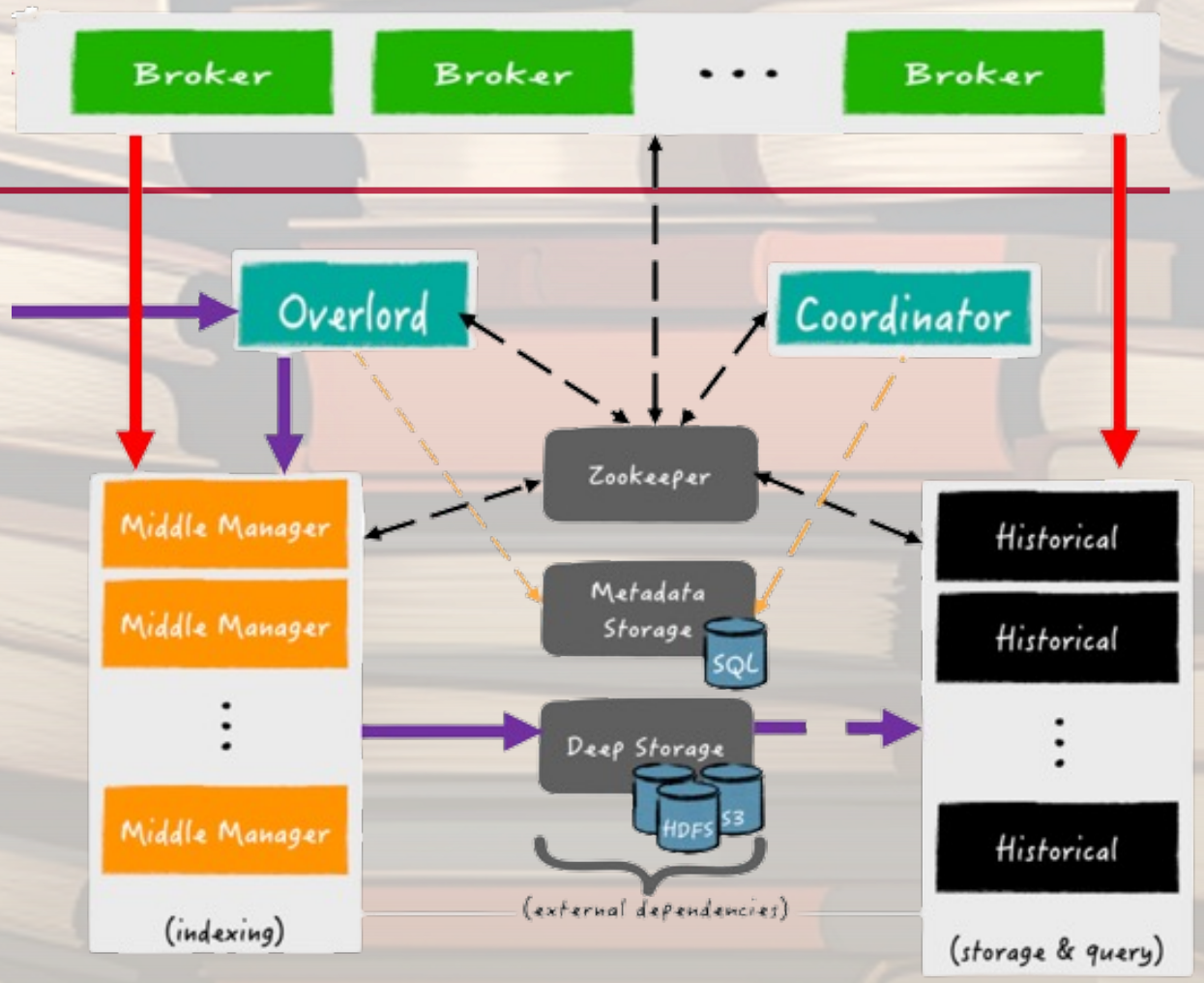
Customer N

Too many tasks

Customer 3

Customer 2

Customer 1



DEPLOY FOR ALL…BOOOM!

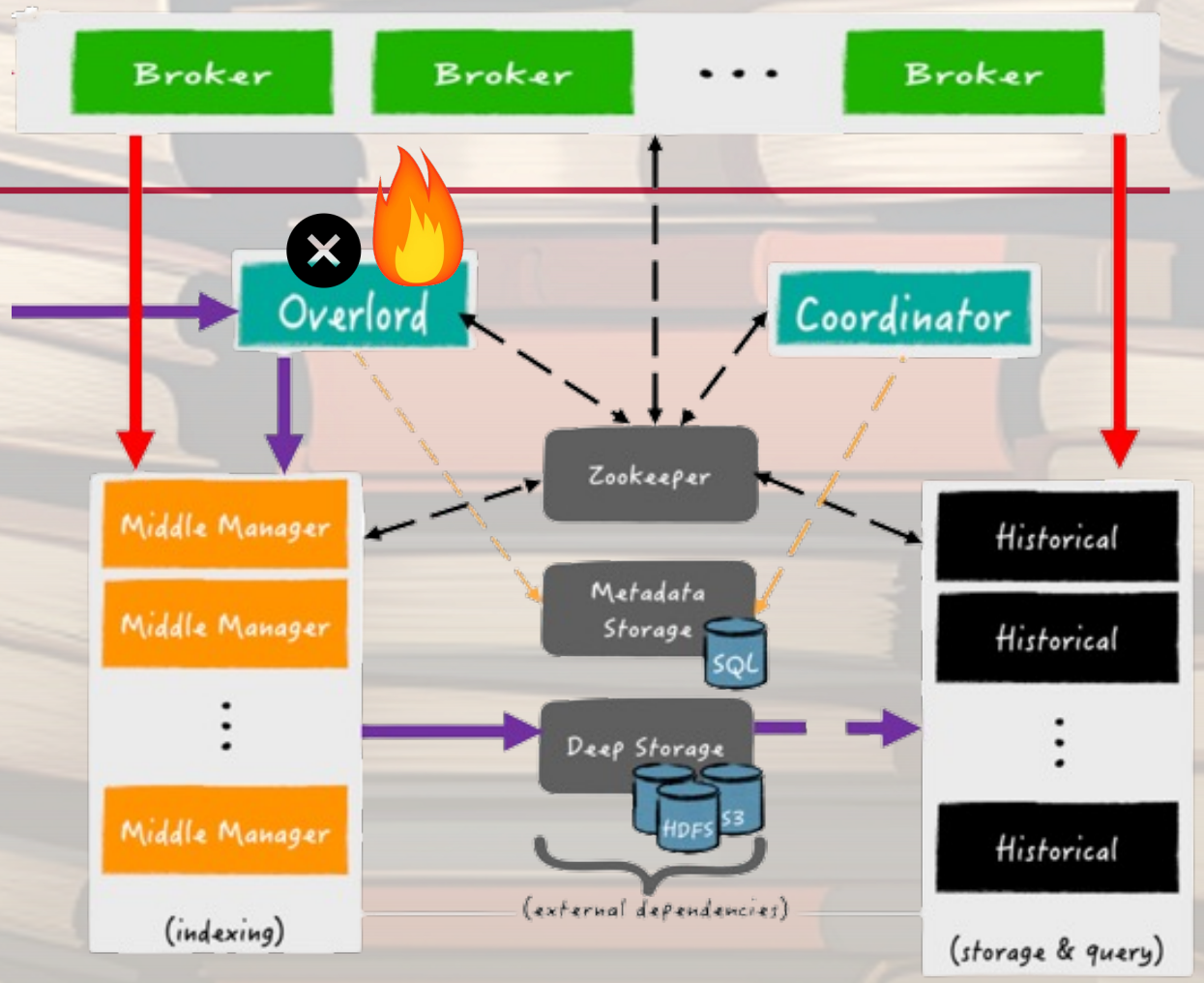
Customer N

Too many tasks

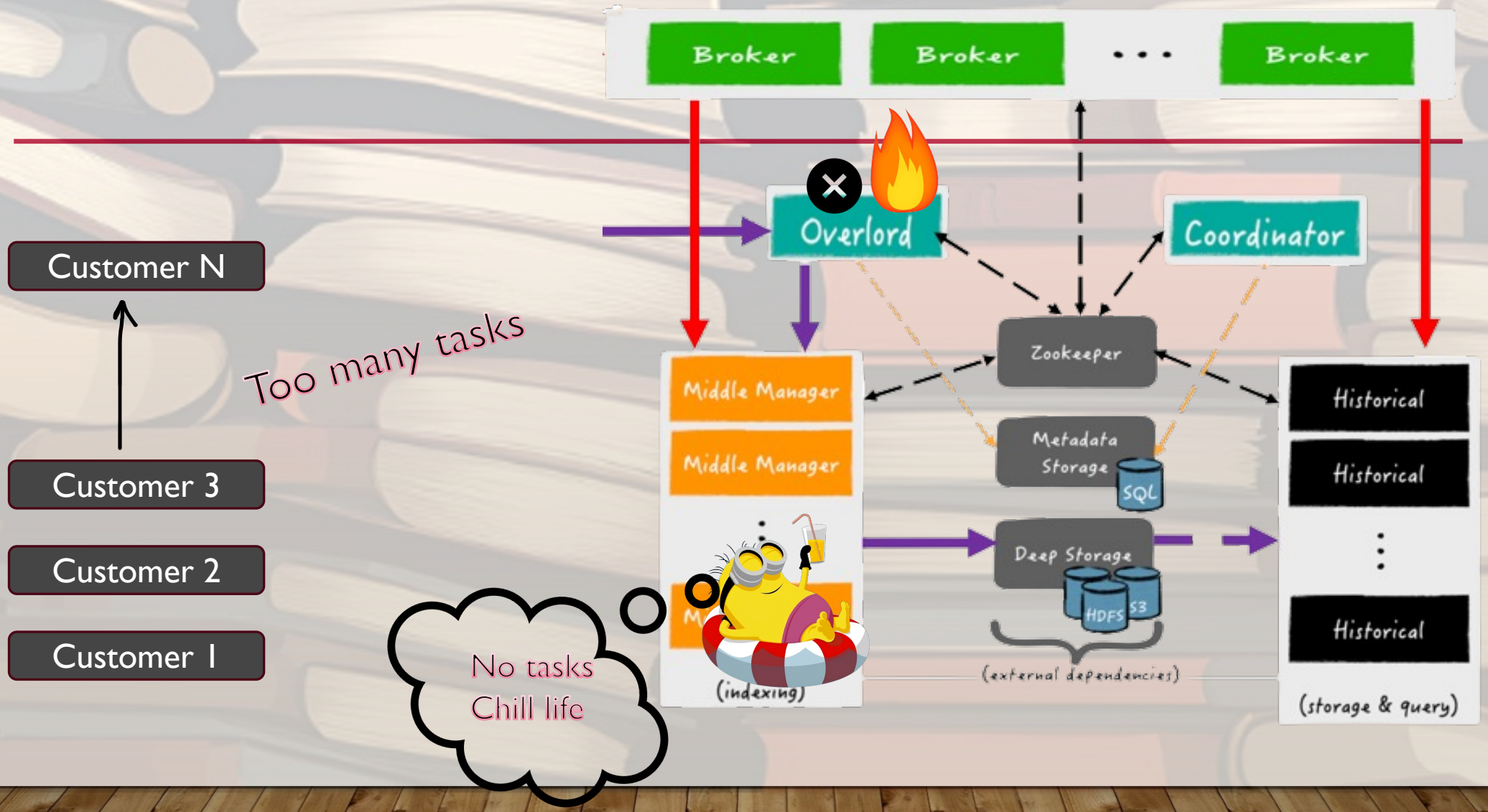
Customer 3

Customer 2

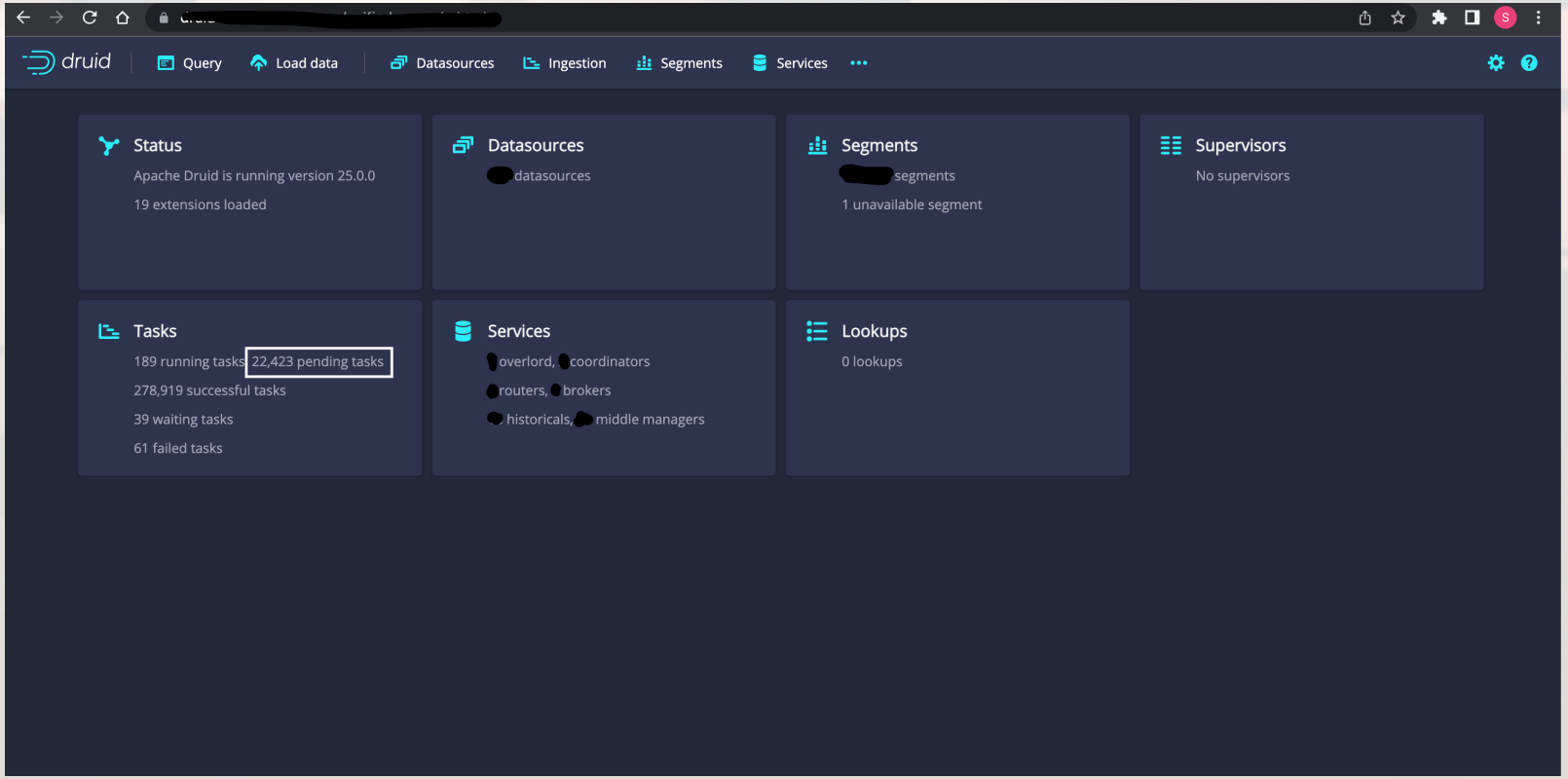
Customer 1



DEPLOY FOR ALL…BOOOM!

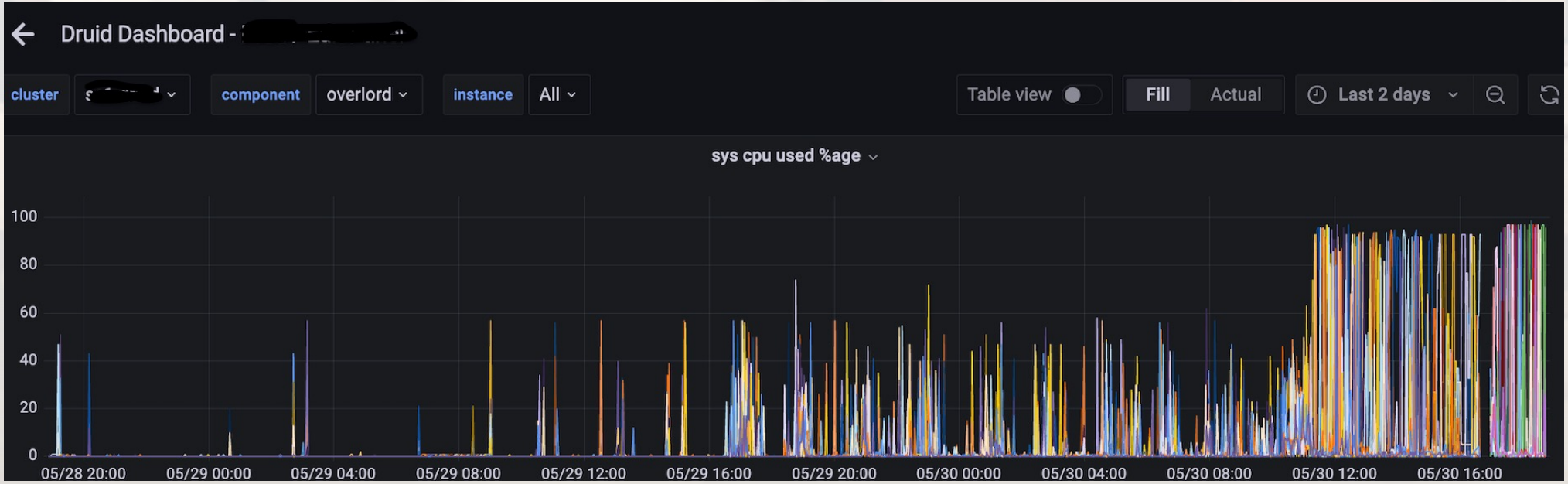


PROOF OF THE PUDDING!



PROOF OF THE PUDDING! 2

PROOF OF THE PUDDING! 3



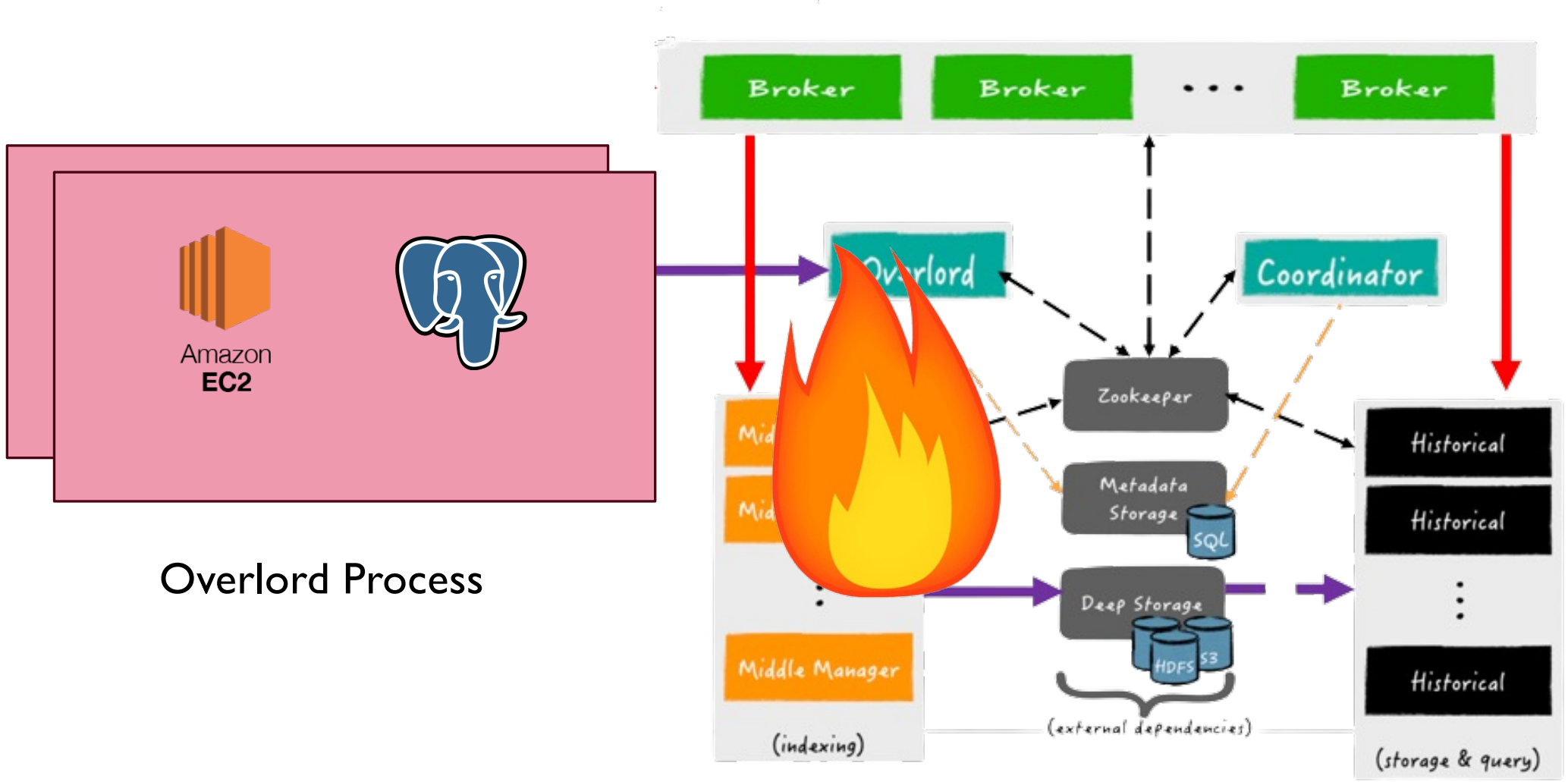
SUMMARY: DRUID INGESTION STRUGGLING

- Ingested smaller, but more tasks.
- Onboarding a few large customers, all good for a day!
- More confidence ☺
- Onboarded all customers at once
  - Ingest task queue kept piling up (till 25K), overwhelmed after 5K
  - Soon, overlord (ingestion orchestrator) node CPU usage at 100%
- All the tasks stuck in pending state
- Task count was 12x more than previous but smaller.
- Ingest tier nodes (MM) were sitting idle, no incoming tasks.
- Ingestion task state not updating as overlord was overwhelmed.

Druid Overlord

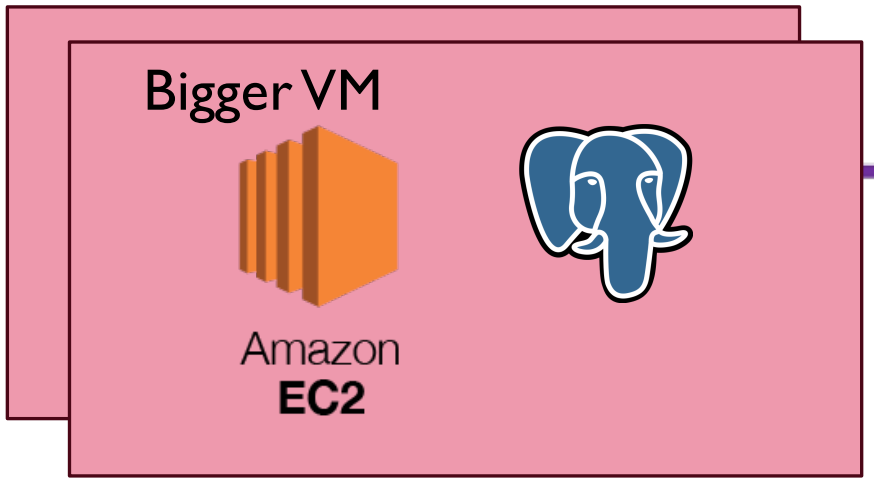
GET INGESTION ORCHESTRATOR ALIVE

INGEST TASKS NOT BEING ASSIGNED!!

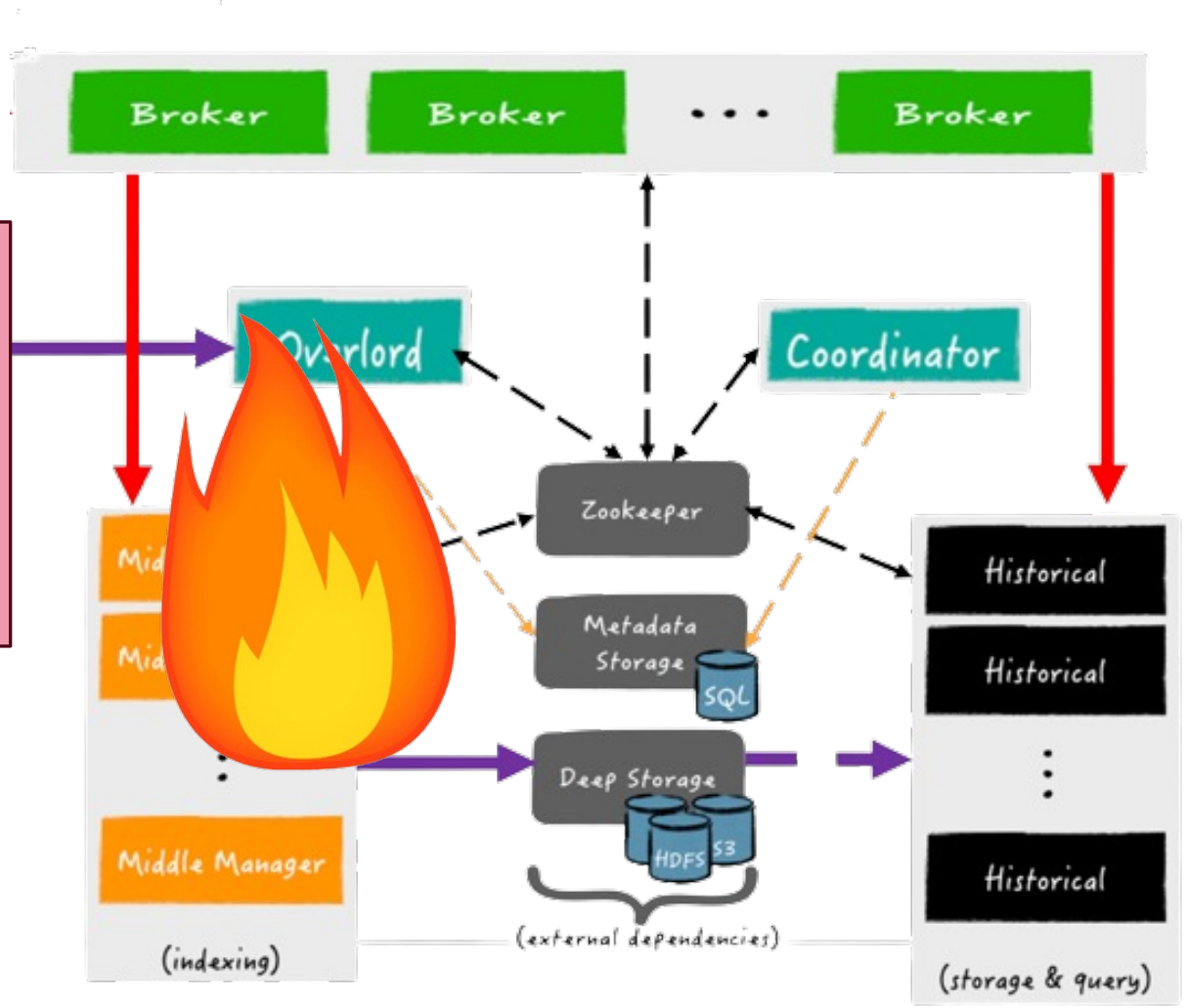


Overlord Process

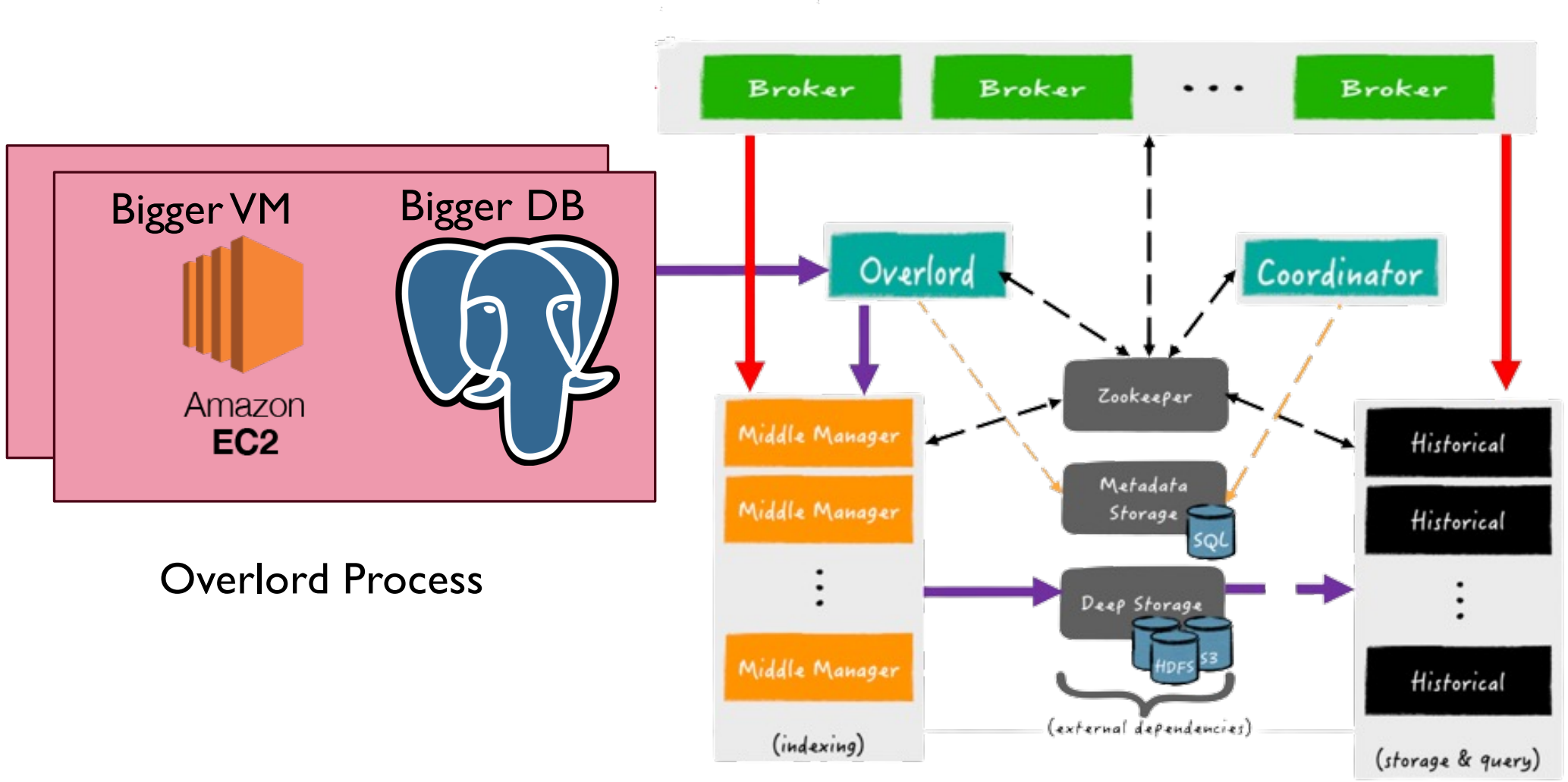
SCALE UP VM (AWS EC2)



Overlord Process



ALSO, SCALE UP DB



Overlord Process

HANDLING THE OVERLORD…

- No support for horizontal scalability
- Only active/passive.
- Vertically scale overlord instance as well as its Postgres DB capacity.
- Optimize the Changed Druid process configs to optimize task assigment!

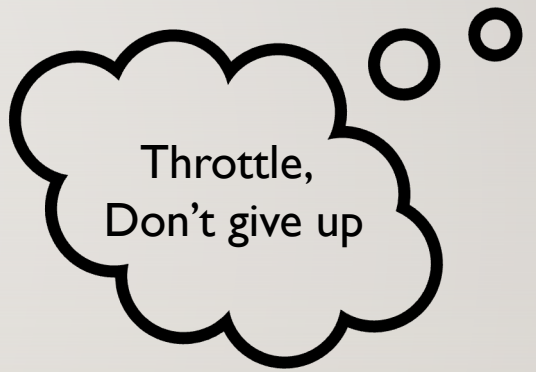
HANDLING THE OVERLORD…

- Optimize the Changed Druid process configs to optimize task assigment!
  - Restrict Global task queue & Throttle!

druid.indexer.queue.maxSize : 5000

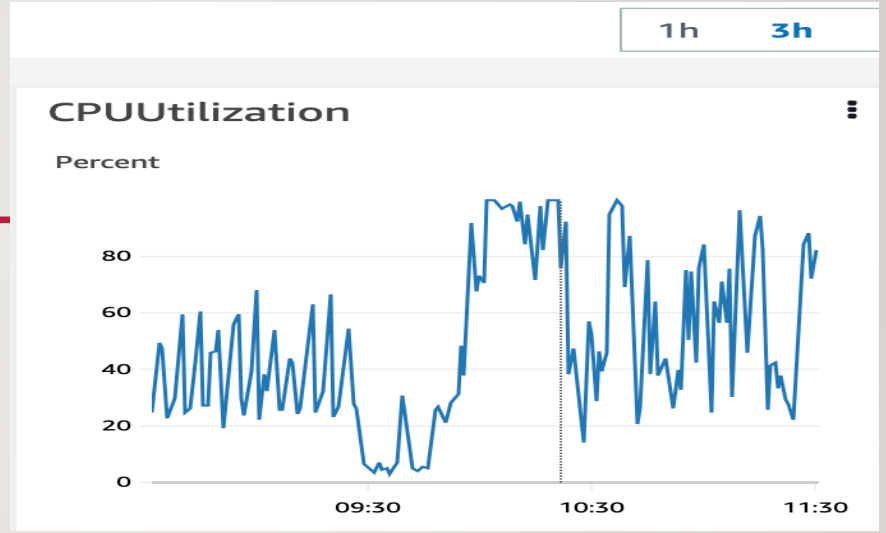
  - Set max pending tasks per customer to 1

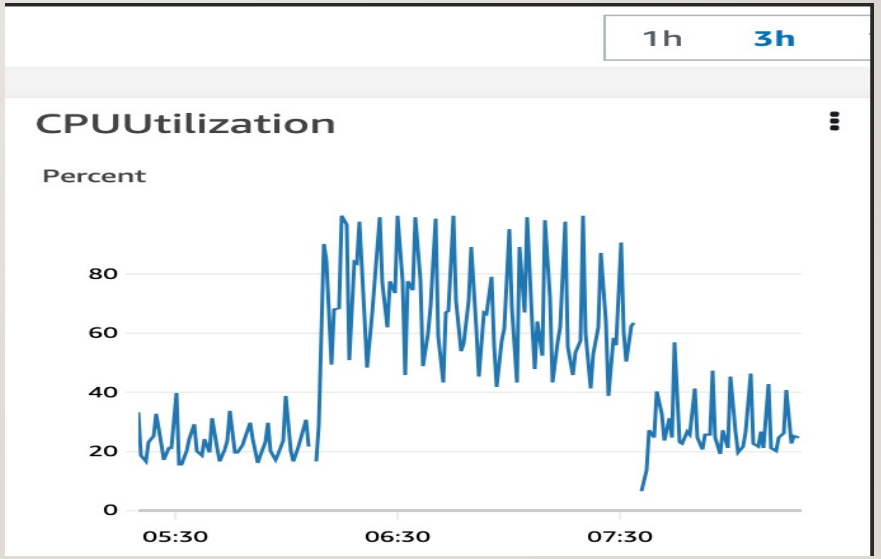
GET /druid/indexer/v1/pendingTasks?datasource=ds1



SCALE UP POSTGRES DB

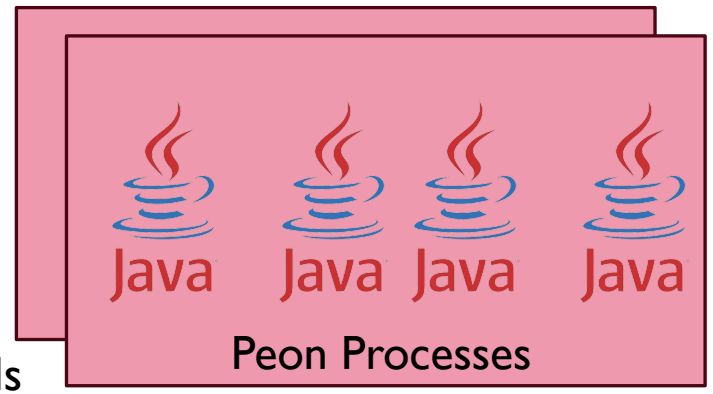
- Queries to Postgres for task were taking long time.
- Add more CPU to DB server
  - Overlord CPU utilization is less
  - Number of pending tasks are less



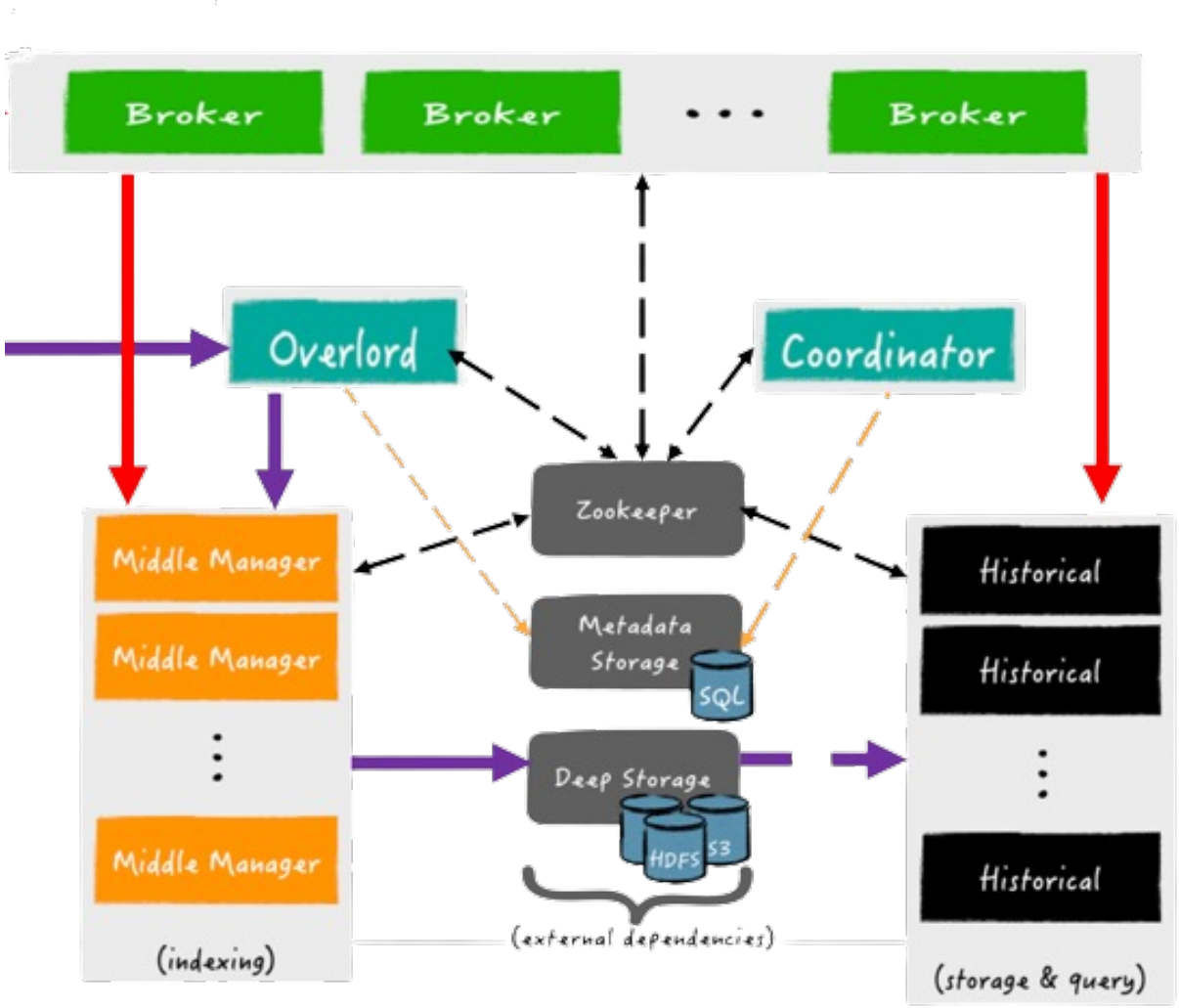


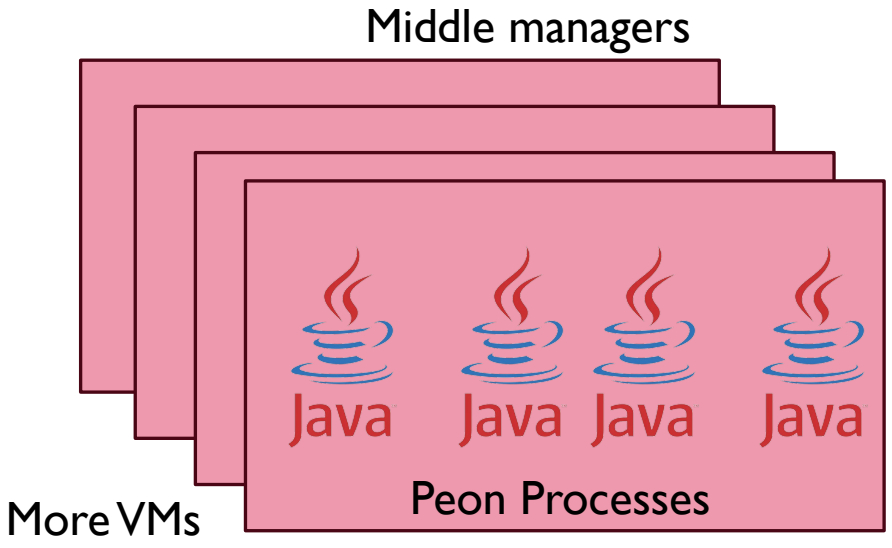
SCALING INGESTION TIER

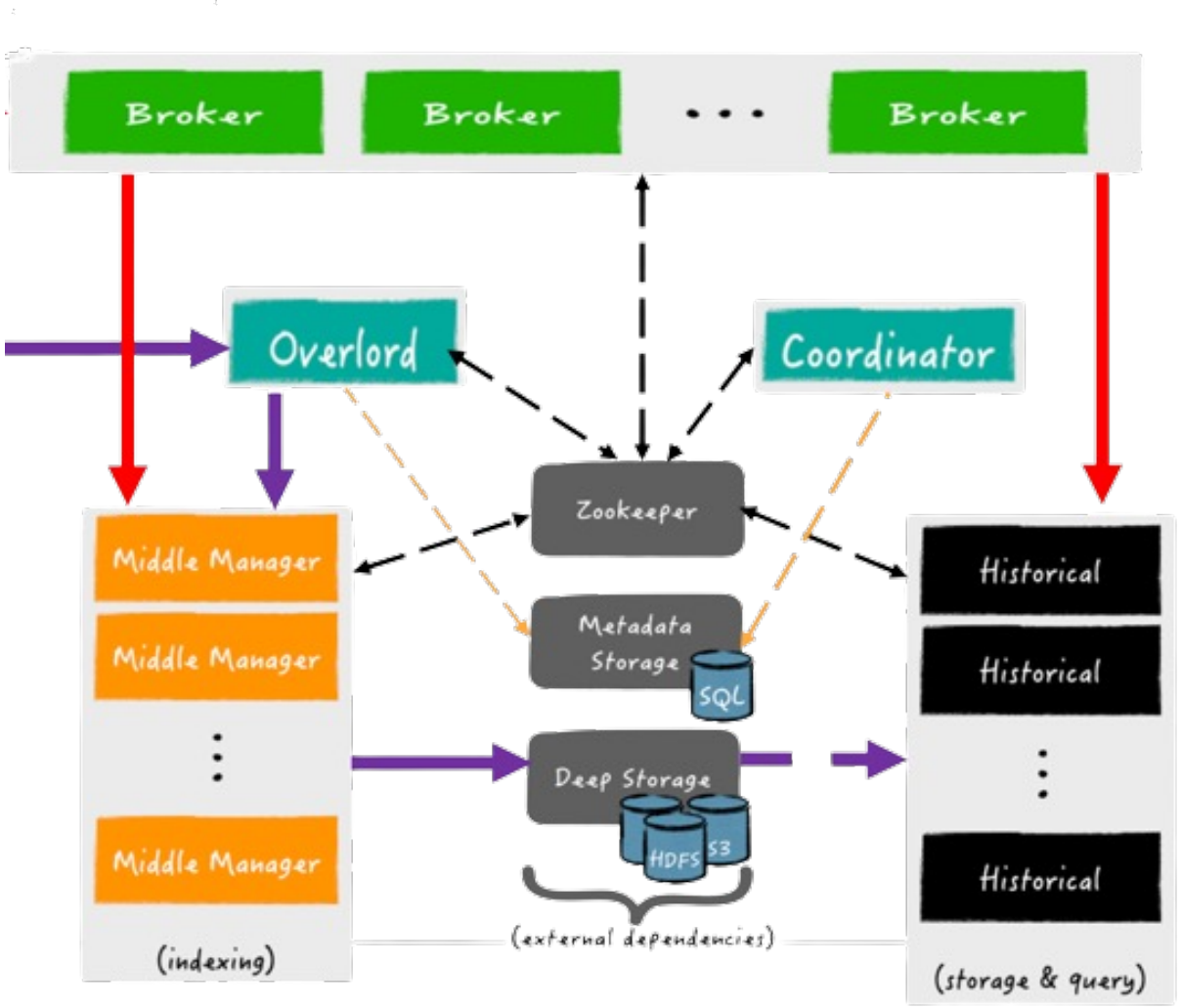
Middle managers



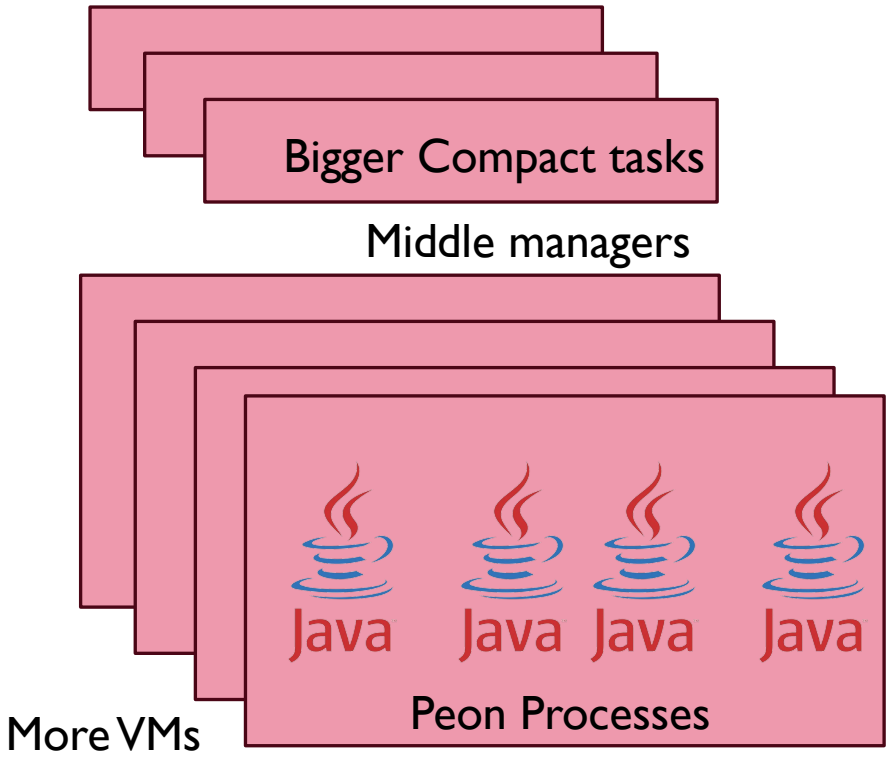
More VMs

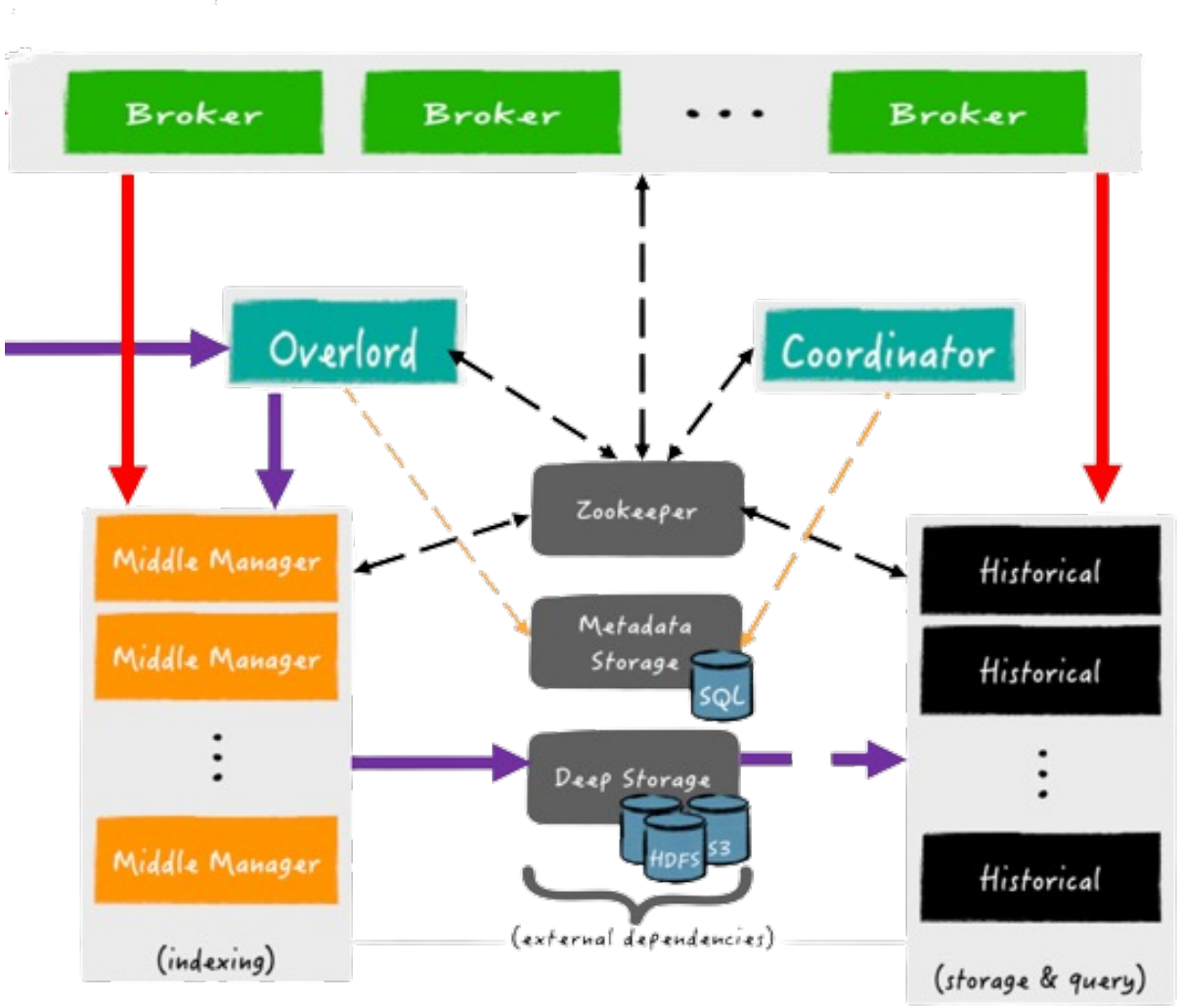



Middle managers
Java
Java
Java
Java
Peon Processes
More VMs

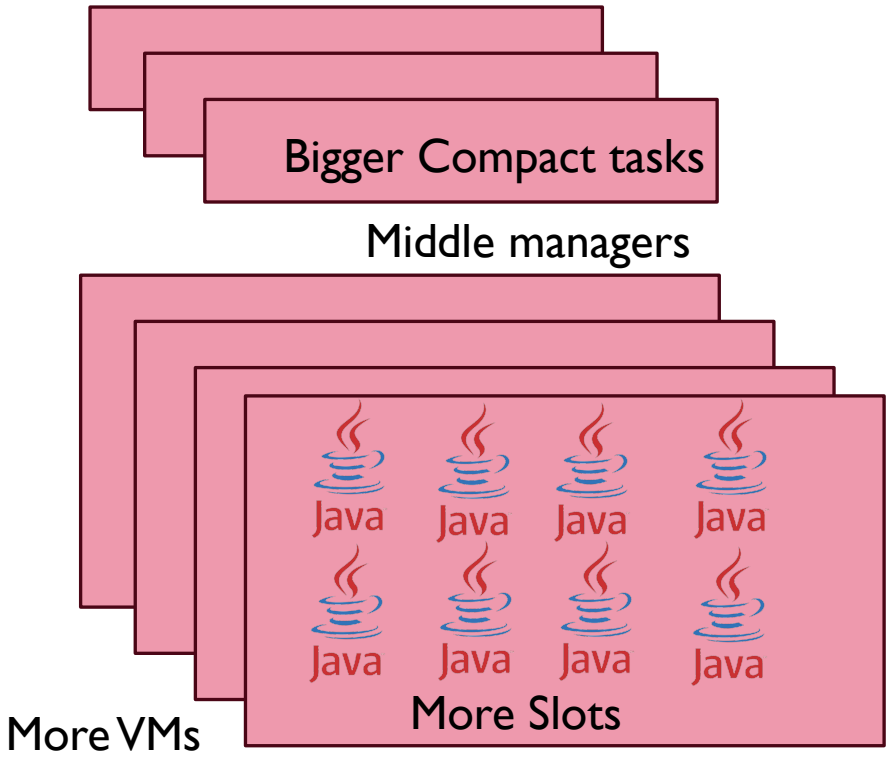

Broker
Broker
Broker
Overlord
Coordinator
Zookeeper
Middle Manager
Middle Manager
Middle Manager
(indexing)
Metadata Storage
SQL
Deep Storage
HDFS
S3
(external dependencies)
Historical
Historical
Historical
(storage & query)

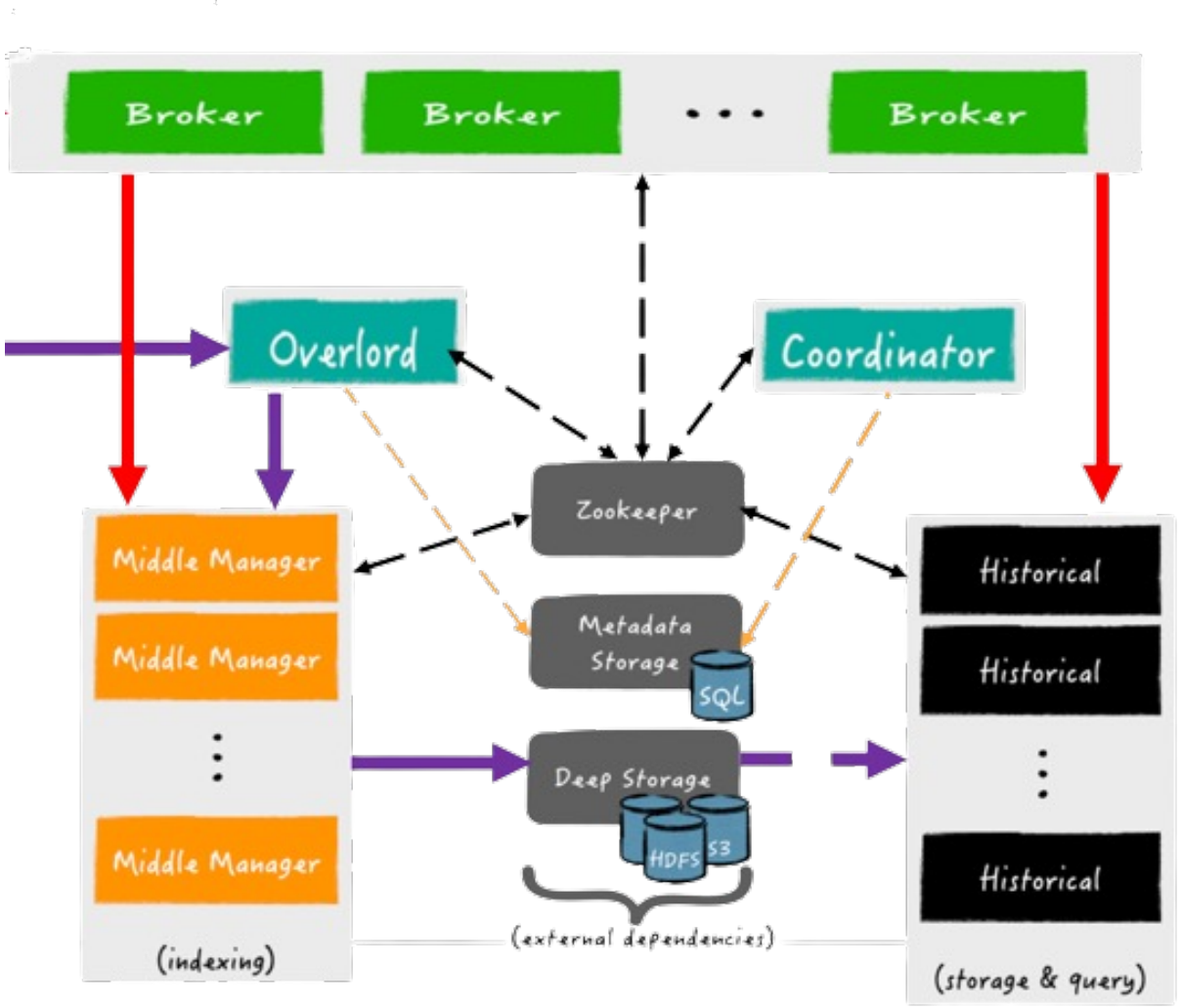
Tiering Middle Managers



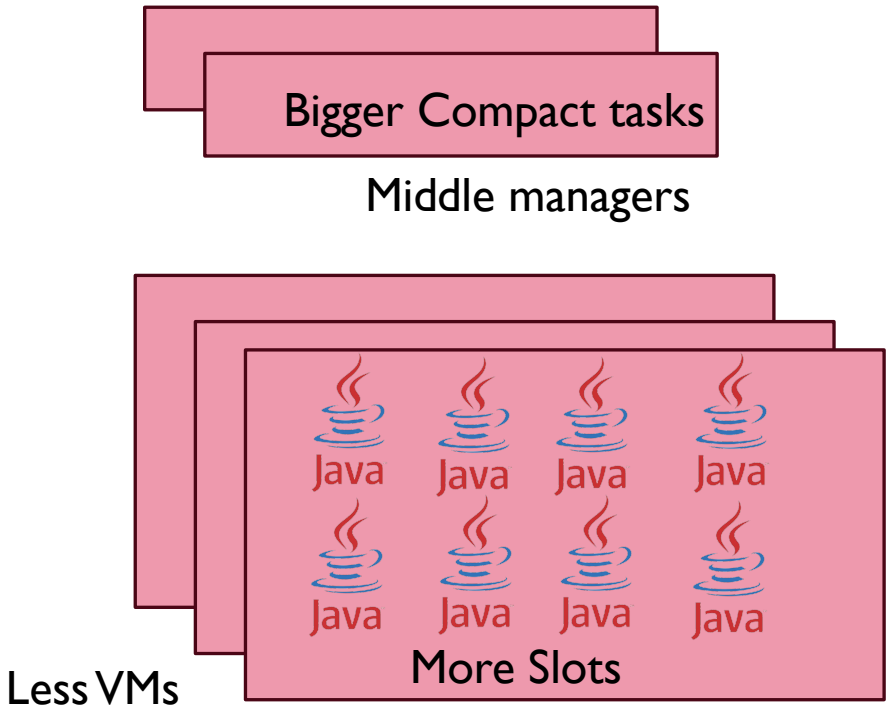


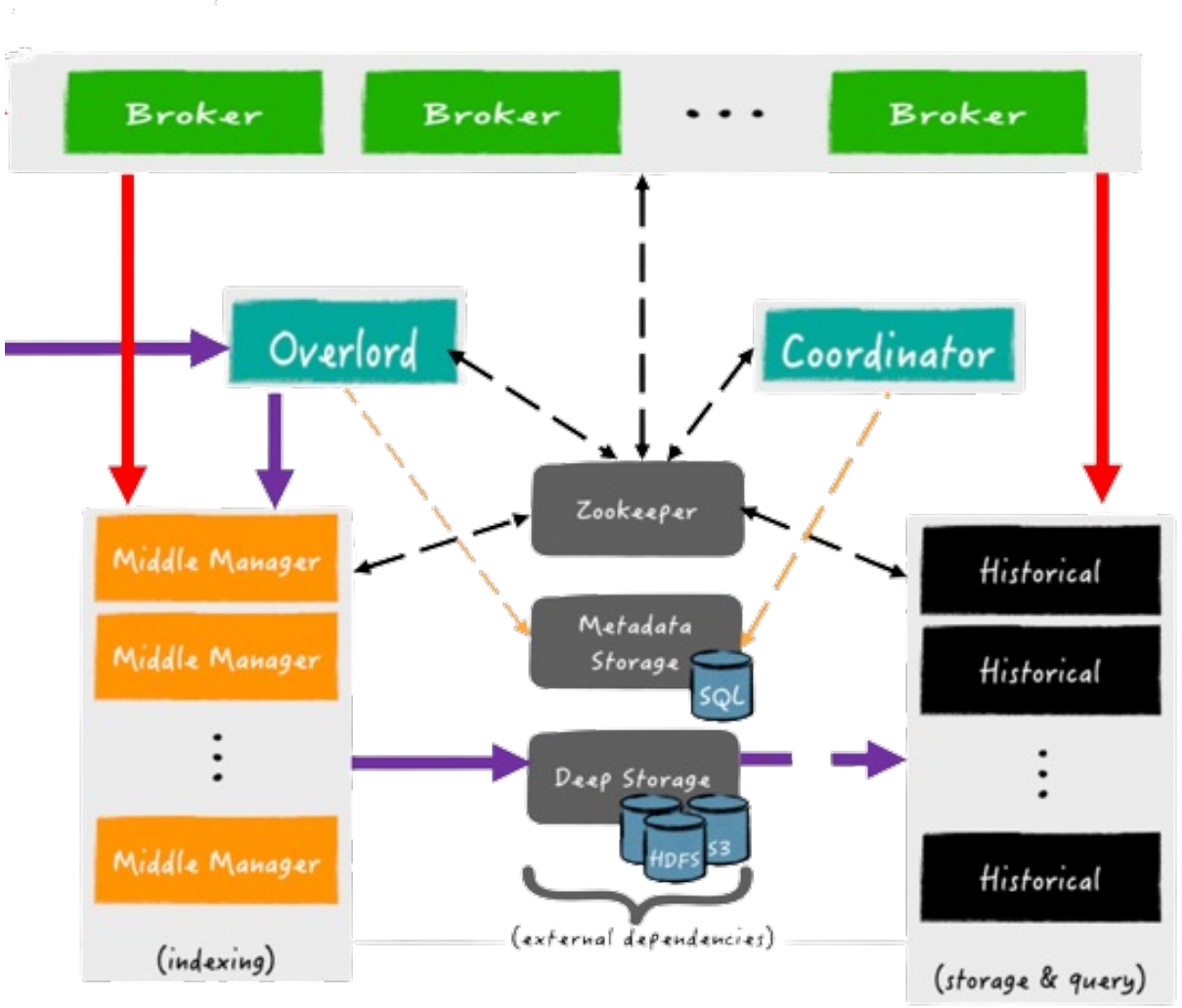
More slots per Middle Managers





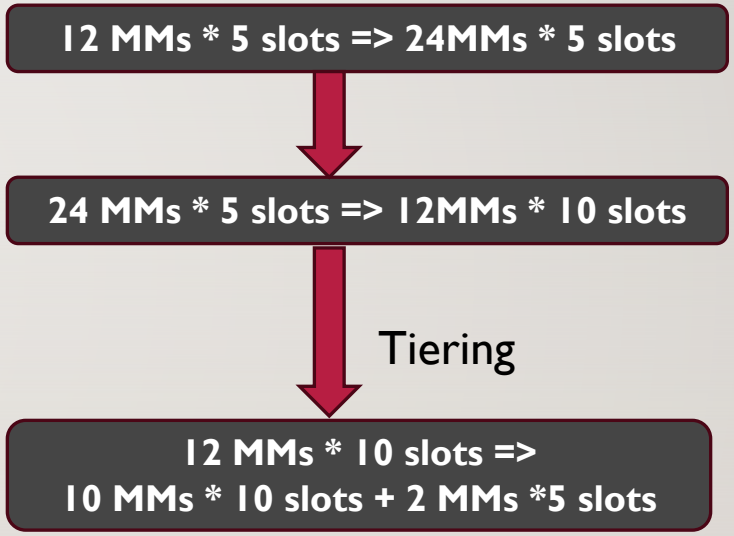
Right size Middle Managers





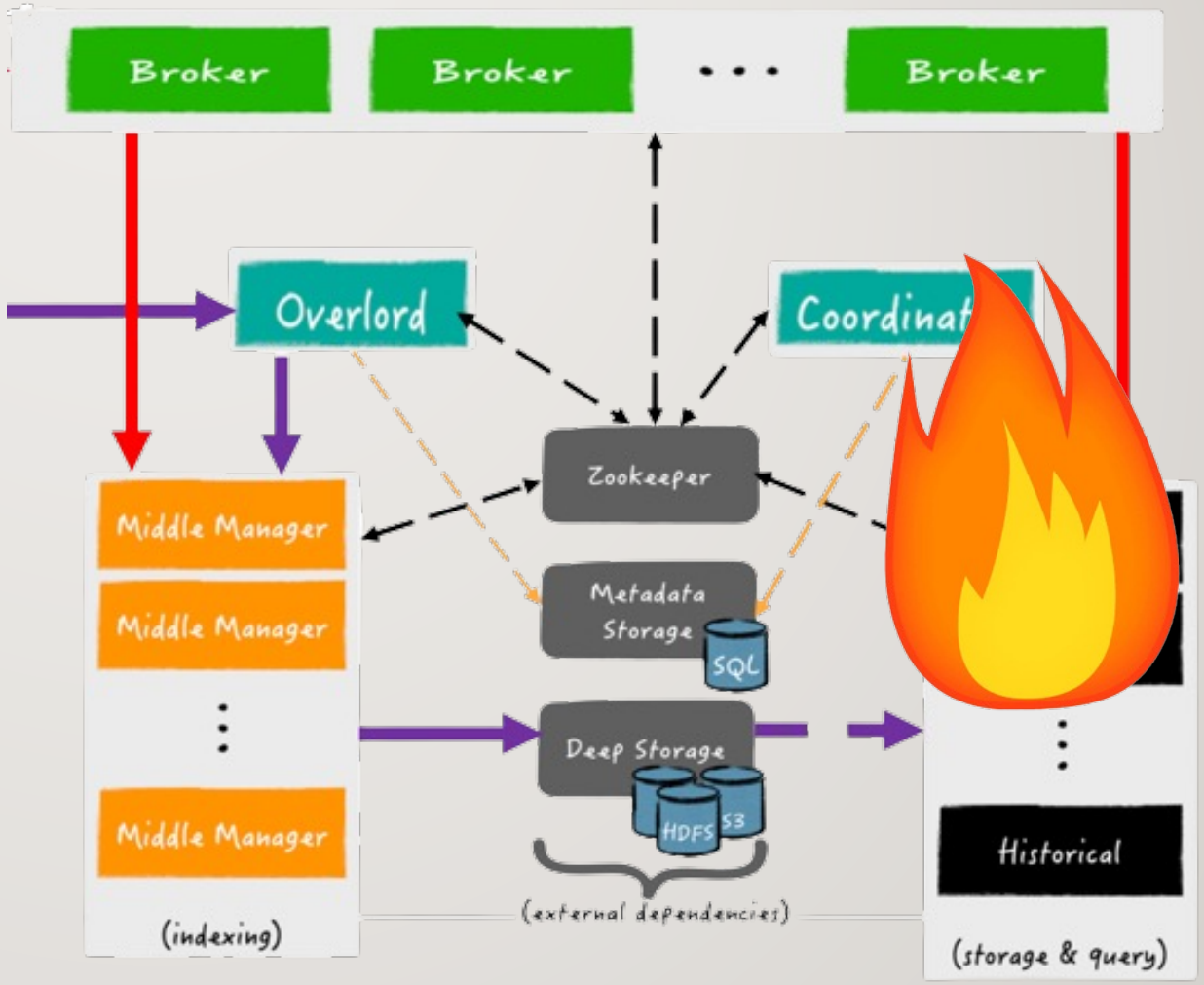
SUMMARY : SCALING MIDDLE MANAGER

- Increased number of middle manager as so that more task slots are available for overlord to assign tasks.
- Then we increased number of slots per middle manager as new tasks were small i.e. having less number of files to ingest.
- We created a separate tier for compaction as these tasks took more resource then the current index tasks.
- Then we right sized the middle manager count in each tier by reducing it.



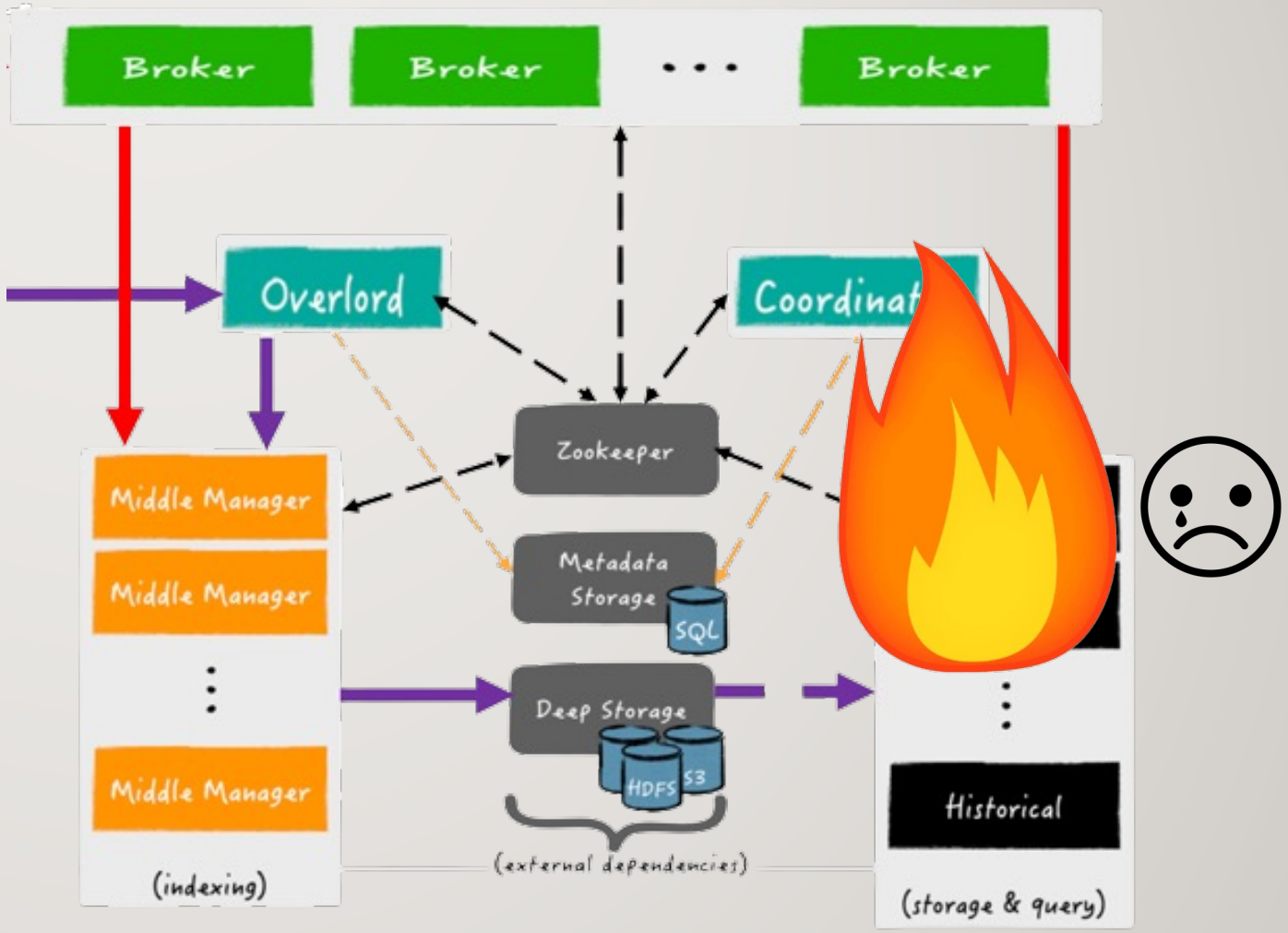
ISSUES IN QUERY TIER

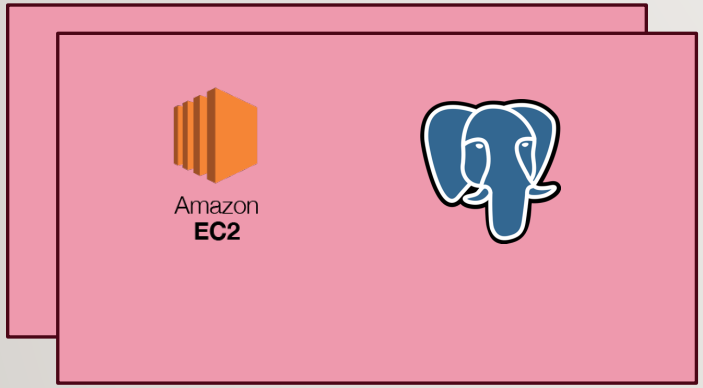
Too many segments to assign !!!



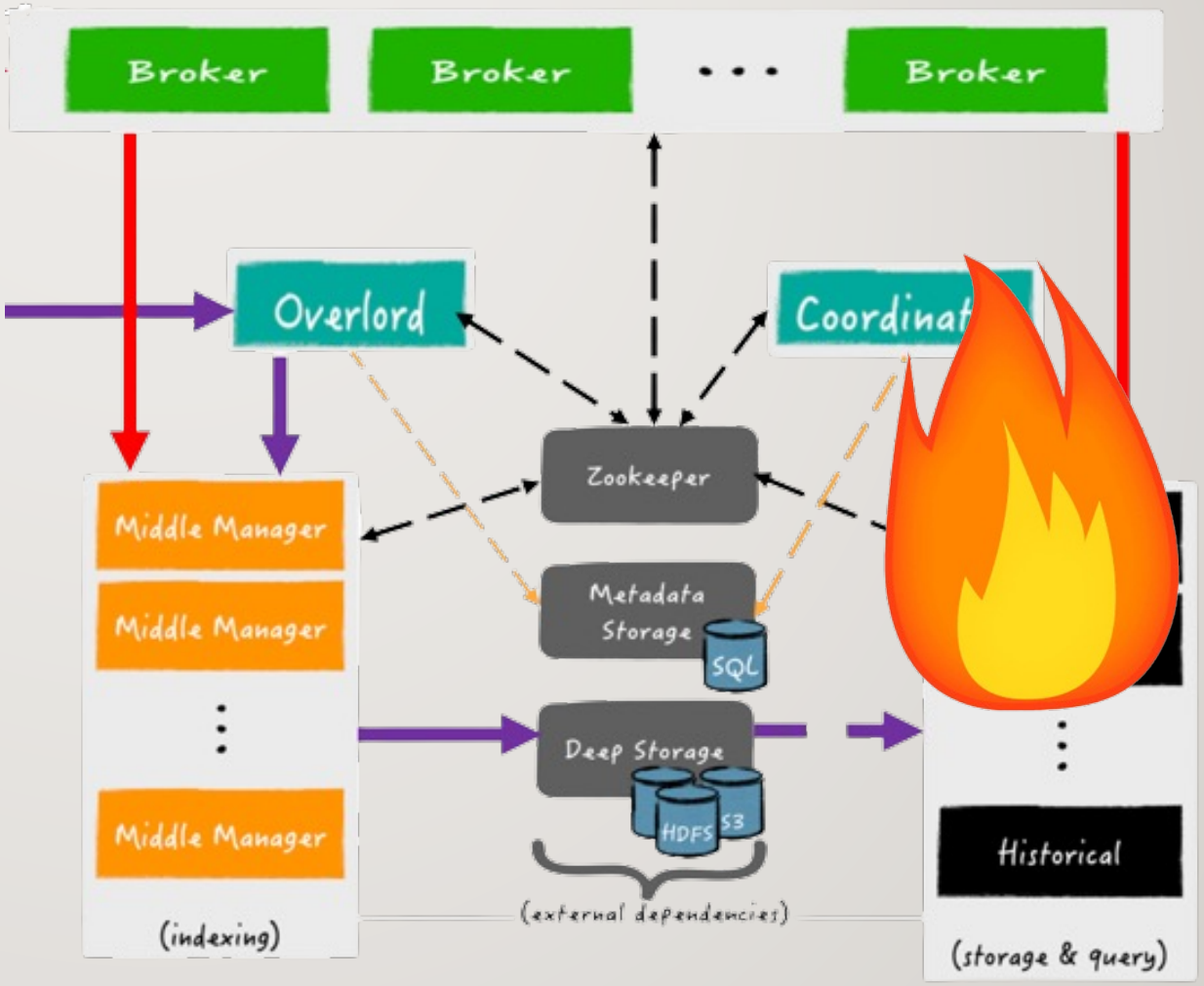
Too many segments to assign !!!

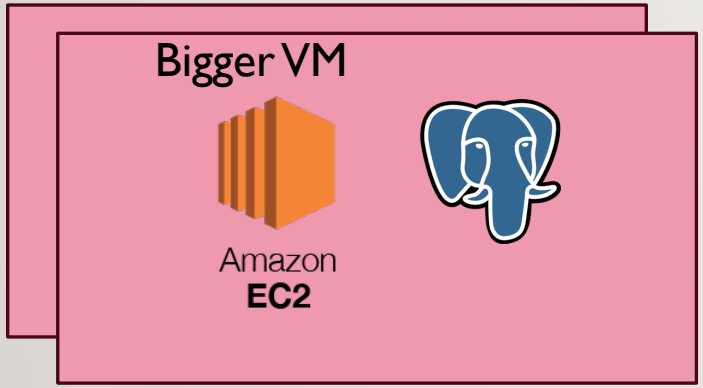
786,830 segments
143,461 unavailable segments



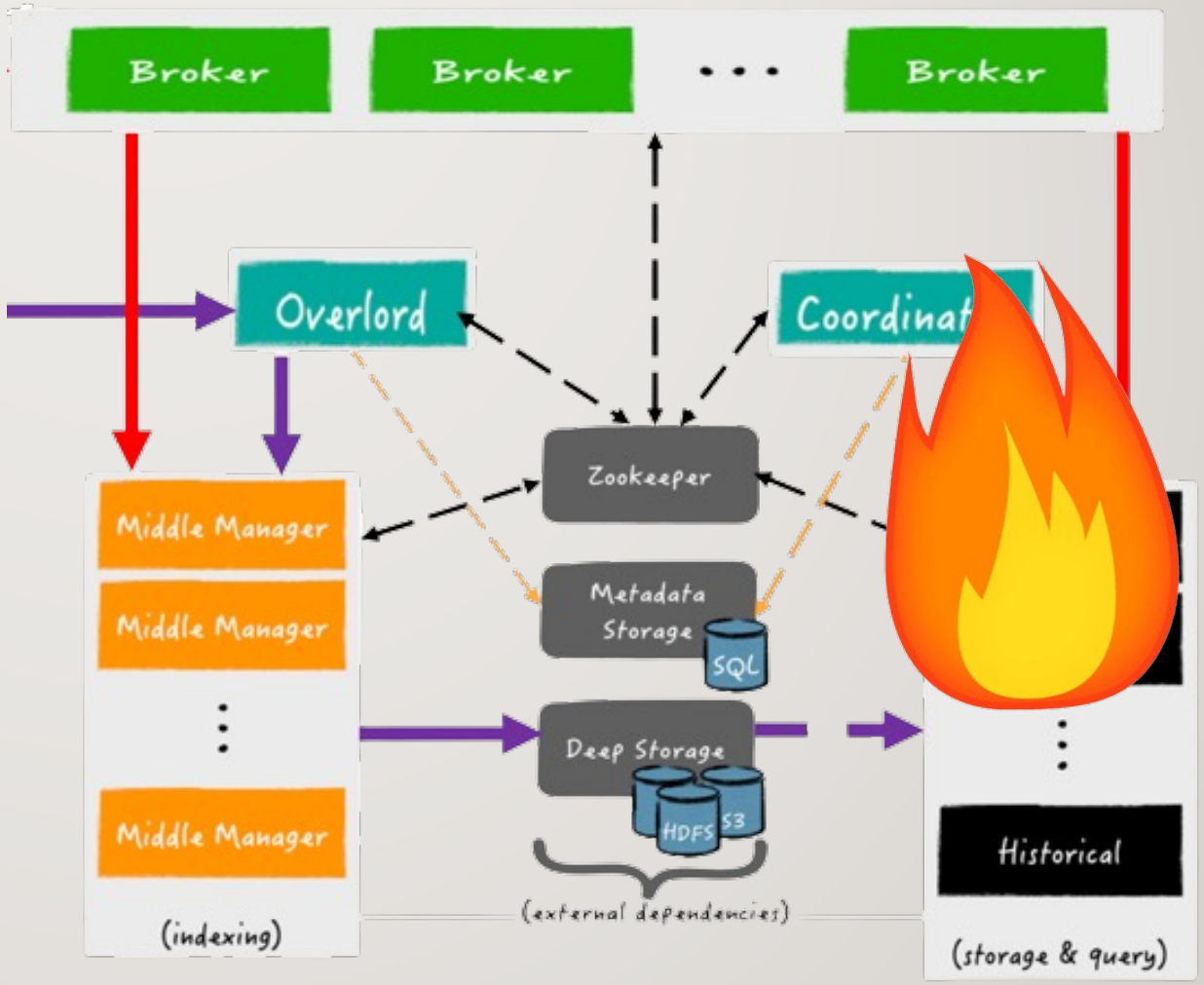


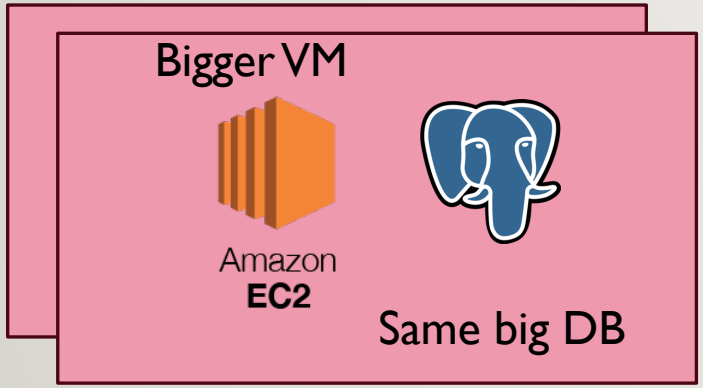
Coordinator Process



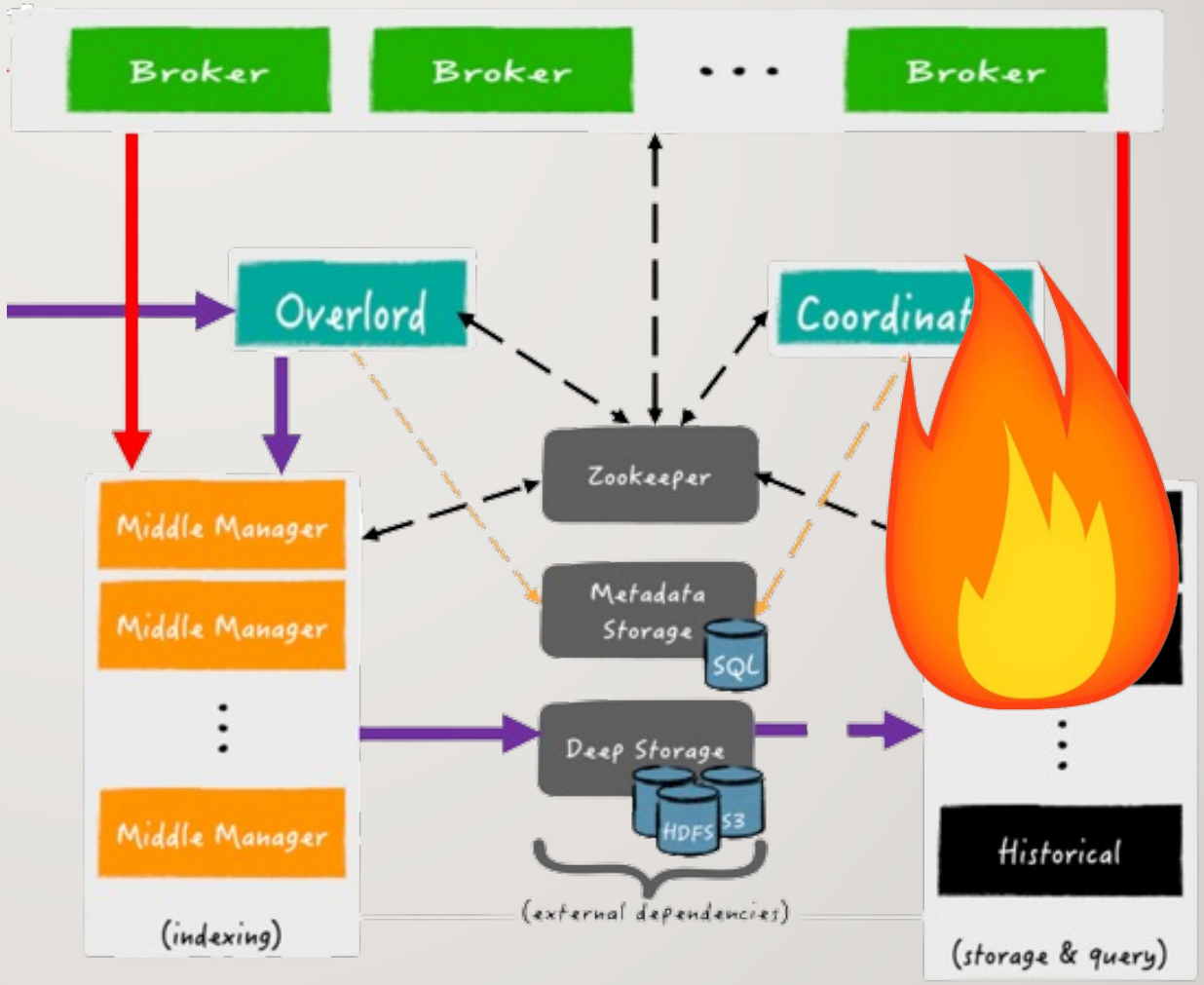


Coordinator Process





Coordinator Process



HANDLING THE COORDINATOR…

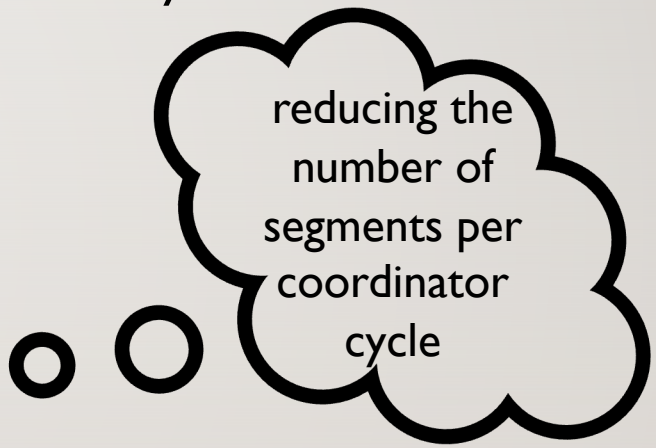
- Increased Coordinator instance type as it is not scalable horizontally
- Tried the following coordinator dynamic configs:

HANDLING THE COORDINATOR…

- Increased Coordinator instance type as it is not scalable horizontally
- Use the **dynamic configs.** Restart on scale is hard!

maxSegmentsToMove: 1000

percentOfSegmentsToConsiderPerMove: 25



HANDLING THE COORDINATOR…

- Increased Coordinator instance type as it is not scalable horizontally
- Tried the following coordinator dynamic configs:

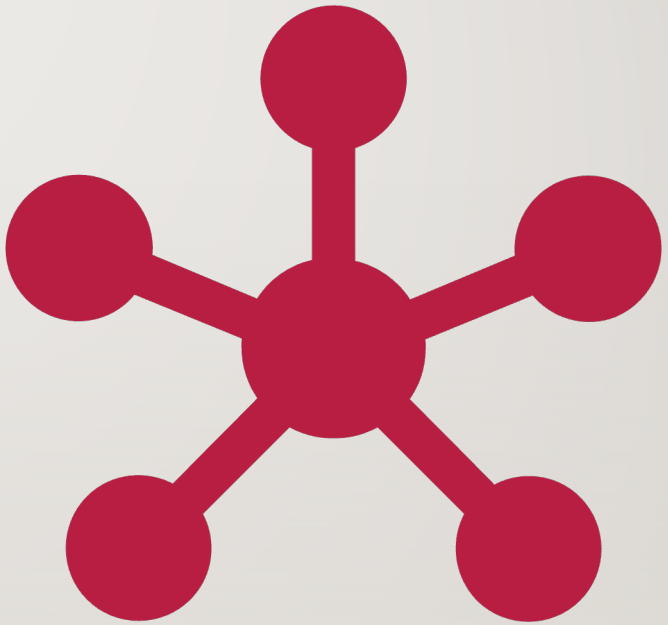
maxSegmentsToMove: 1000

percentOfSegmentsToConsiderPerMove: 25

useRoundRobinSegmentAssignment: true

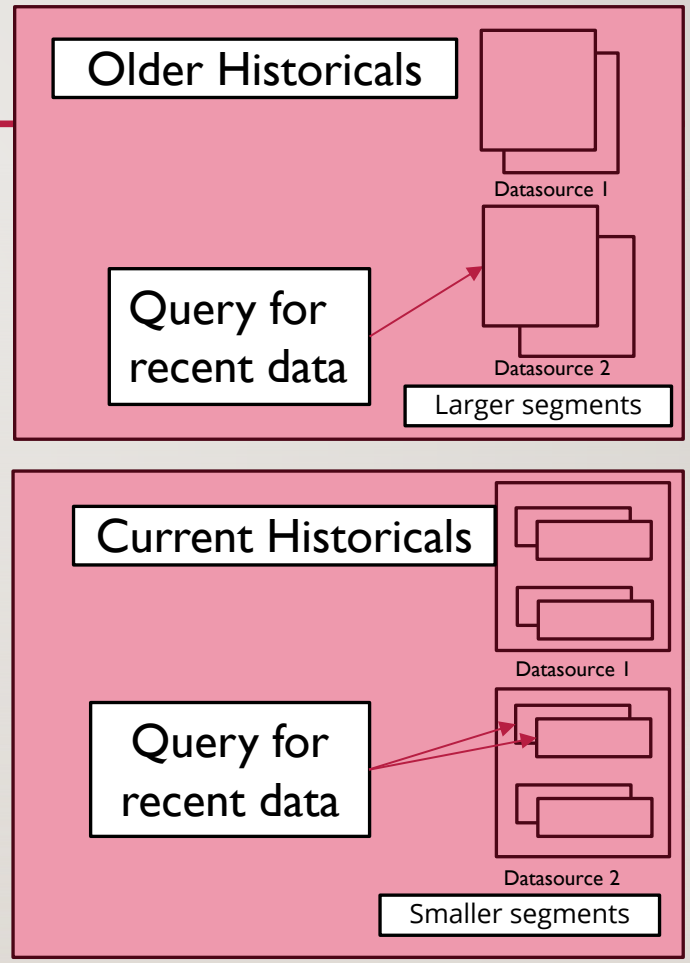
Assign segments In round-robin fashion first. Lazily reassign with chosen balancer strategy later

FINALLY, THE QUERY NODES

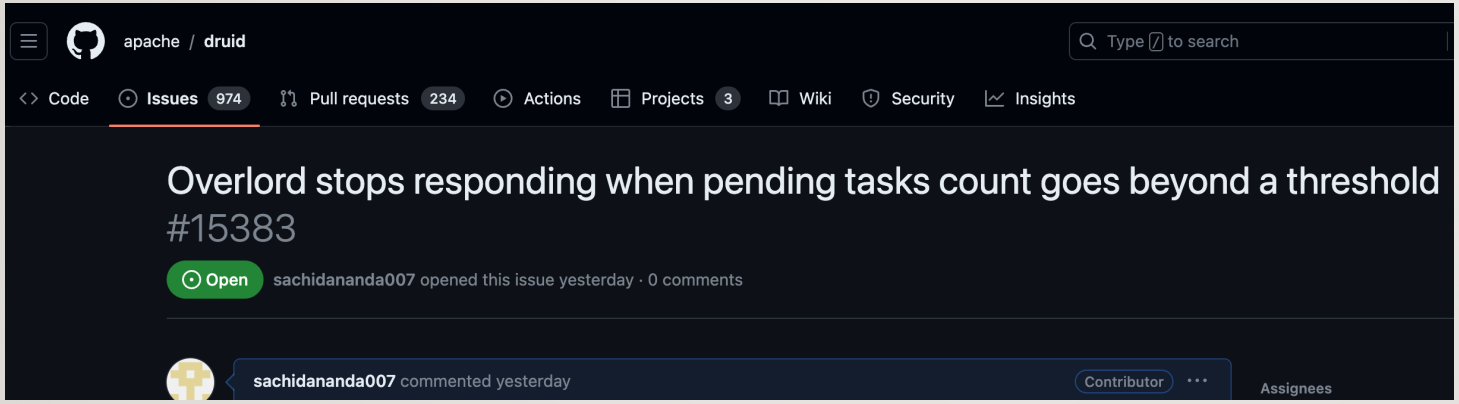


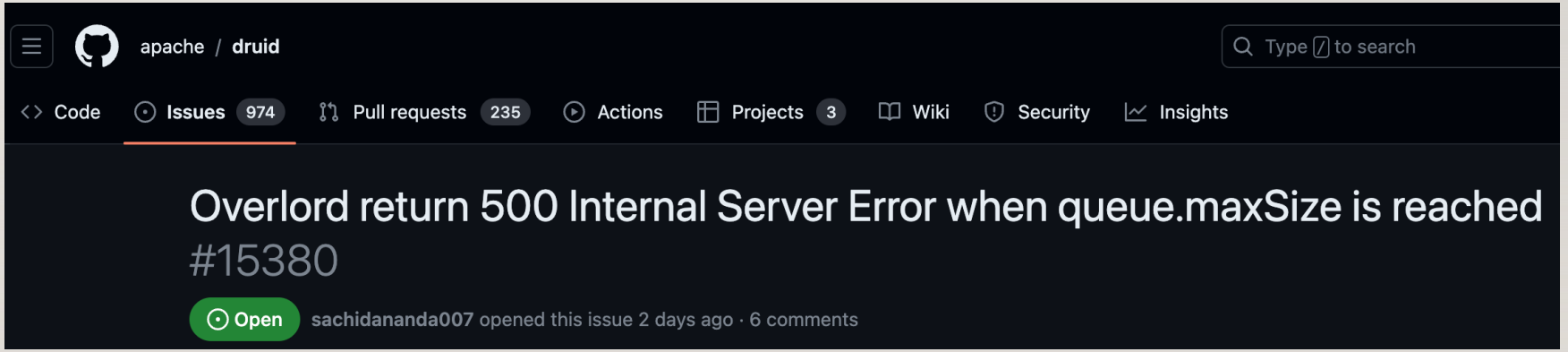
HANDLING THE HISTORICALS

- Until auto compaction done:
  - More no of segments for queries
  - More resources to power query tier
- Beware: More segments = more cache to load
  - Check for linux configs
    - max_map_count

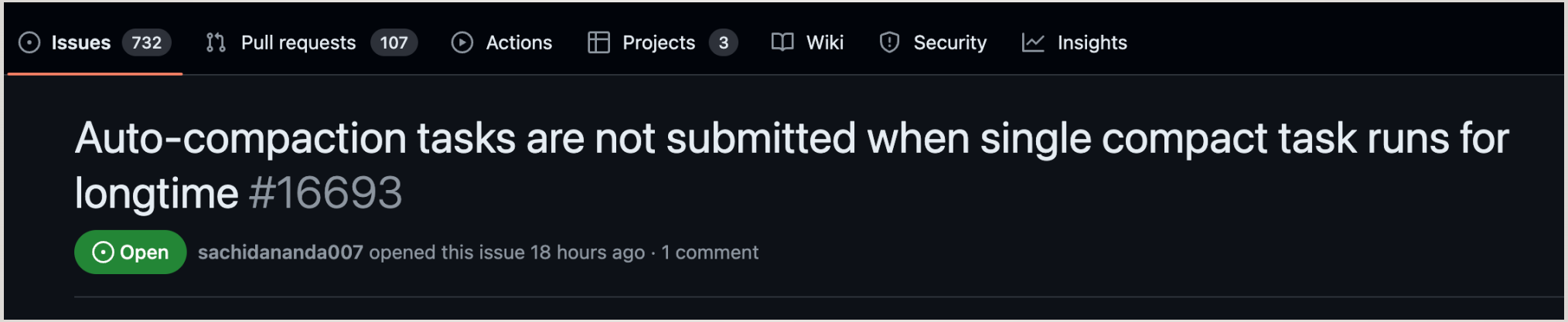


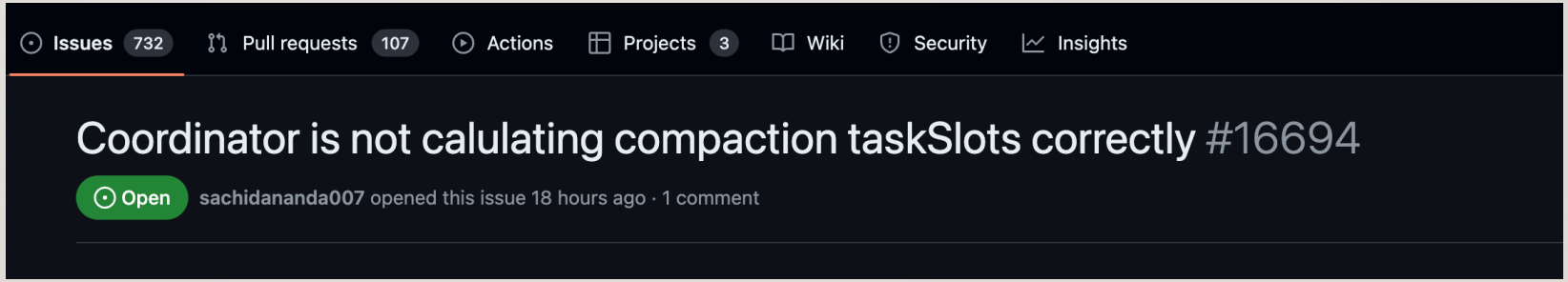
FILE ISSUES IN OSS



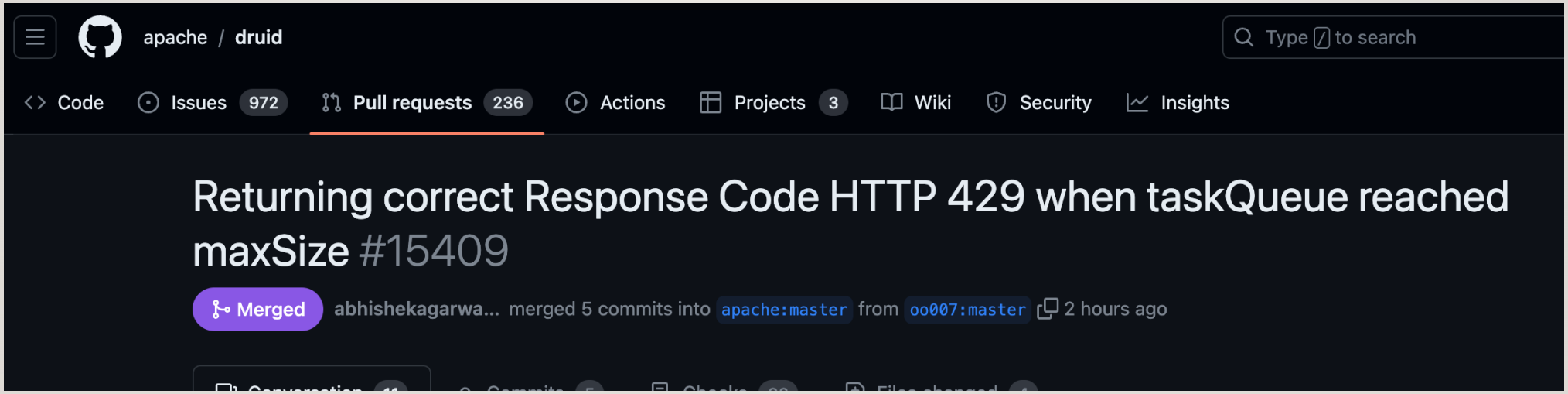


FILE ISSUES IN OSS

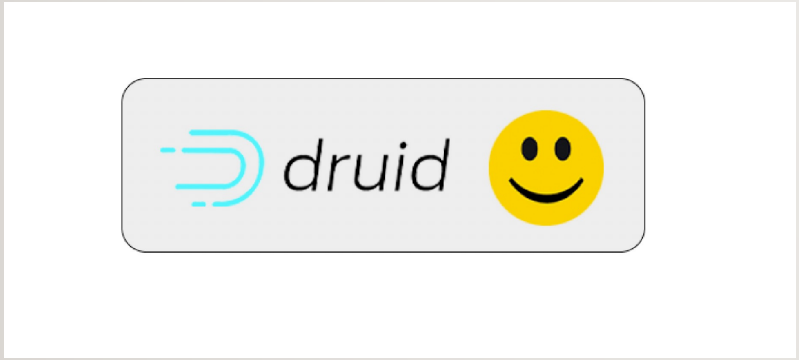


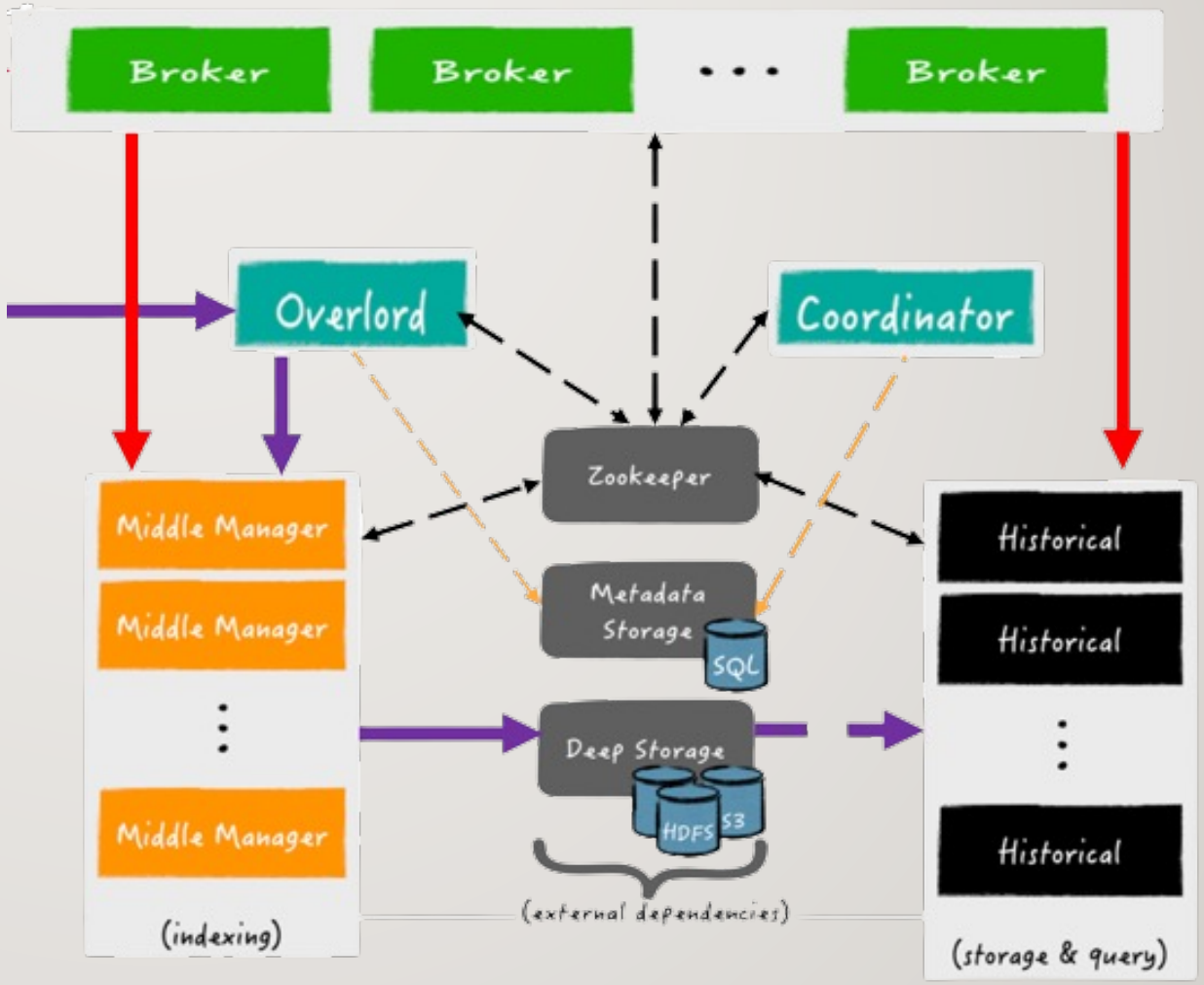


FIX ISSUES TOO IF POSSIBLE!



Happy State!!!


druid


Broker
Broker
Broker
Overlord
Coordinator
Zookeeper
Middle Manager
Middle Manager
Middle Manager
Metadata Storage
SQL
Deep Storage
HDFS
S3
Historical
Historical
Historical
(indexing)
(external dependencies)
(storage & query)

QUESTIONS ?

Liked this?
Checkout our upcoming meetups!

THANK YOU

Shivji Kumar Jha

linkedin.com/in/shivjijha/

slideshare.net/shiv4289/presentations/

youtube.com/@shivjikumarjha

Sachidananda Maharana

https://www.linkedin.com/in/sachidanandamaharana/