HAP: 多流动态实时分析系统

(A multi-streams dynamic real-time analytic system)

张李晔

新氦科技

Helium

he2.io

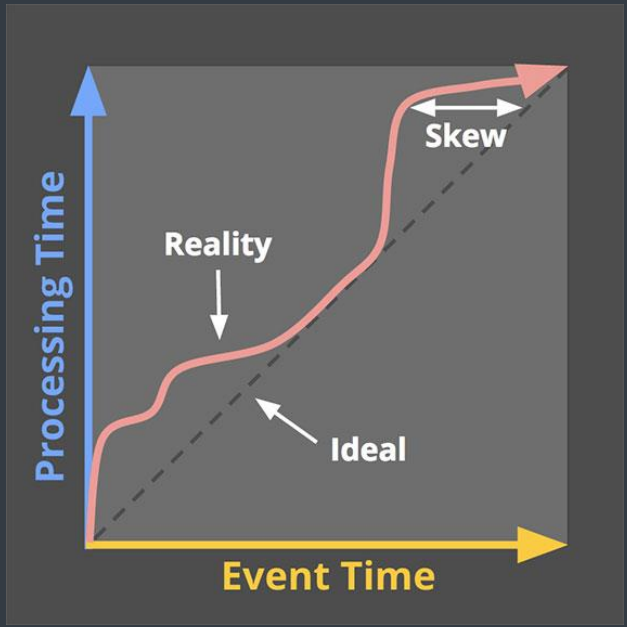
Agenda

- Big data environment
- What is HAP
- HAP Arch
- Streaming underlying
- Case study
- Conclusion

Big data environment

- Batch
- Streaming
- Lambda
- Kappa

4 questions for unbounded data processing*



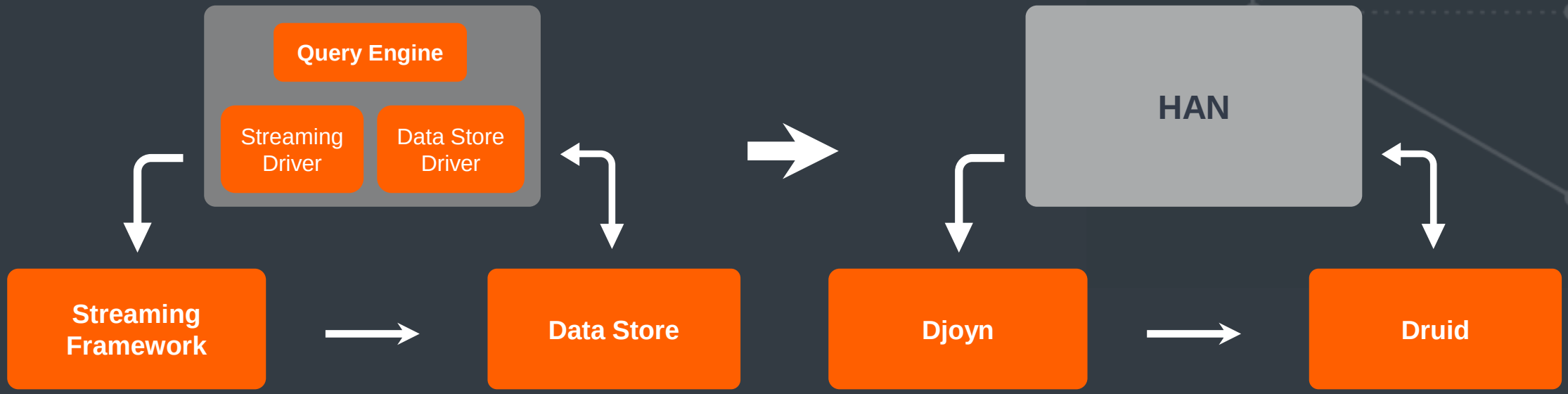
- What results are calculated?
- Where in event time are results calculated?
- When in processing time are results materialized?
- How do refinements of results relate?

* *https://www.oreilly.com/ideas/the-world-beyond-batch-streaming-102*

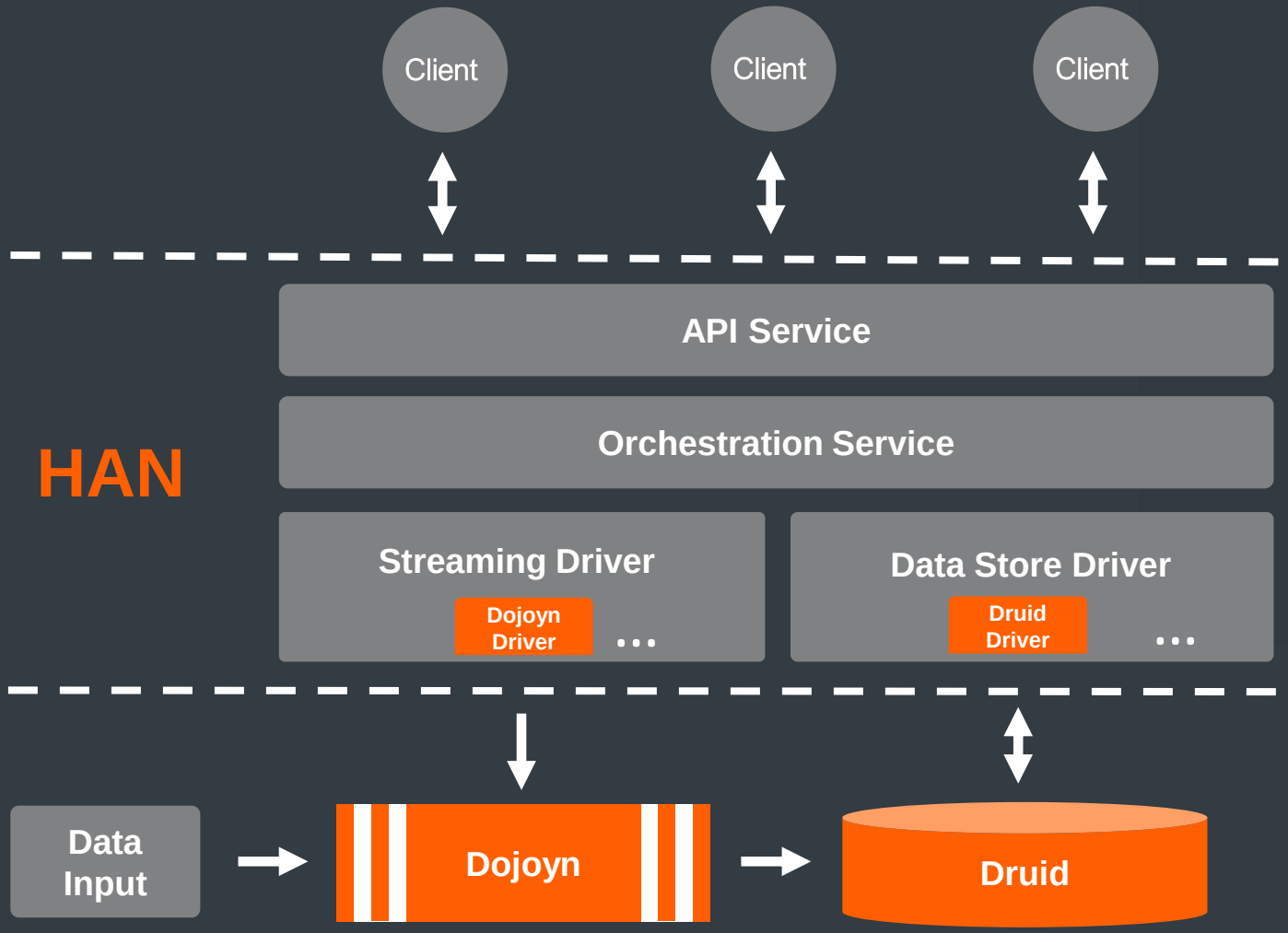
What is HAP

- K8s based software stack
- HAP philosophy
  - One query for everything
  - “Eliminating” window in streaming
  - Multi streaming support
  - Dynamic job creation/deletion
  - Low latency
  - Sub-second analyzing

Computing model



HAP Arch

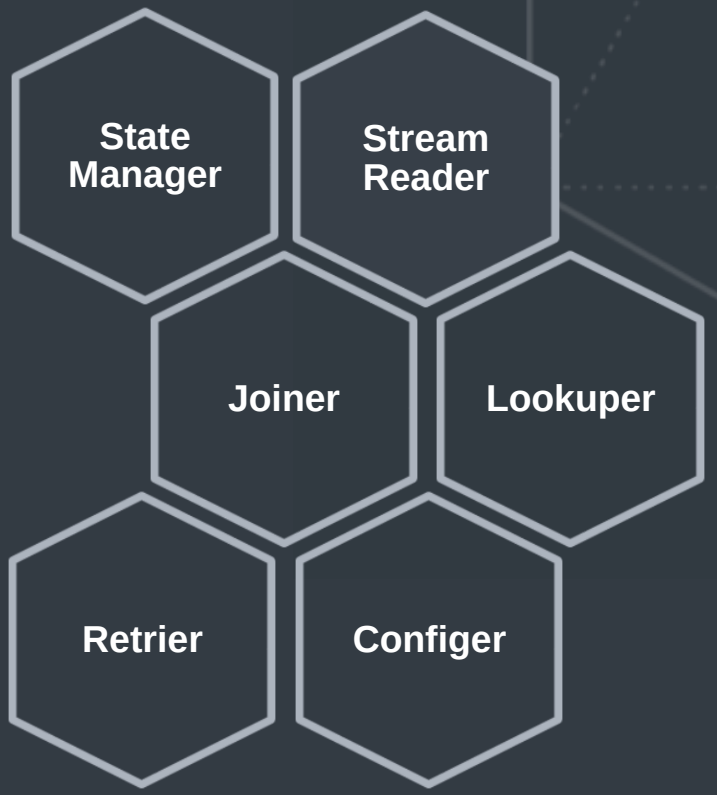

Client
Client
Client
API Service
HAN
Orchestration Service
Streaming Driver
Dojoyn Driver
...
Data Store Driver
Druid Driver
...
Data Input
Dojoyn
Druid

Data store underlying

- Druid
  - Real-time Streaming Ingestion
  - Sub-second OLAP Queries
  - Highly Available
  - Horizontally Scalable
  - Trillions of events, petabytes of data
  - Thousands of queries per second

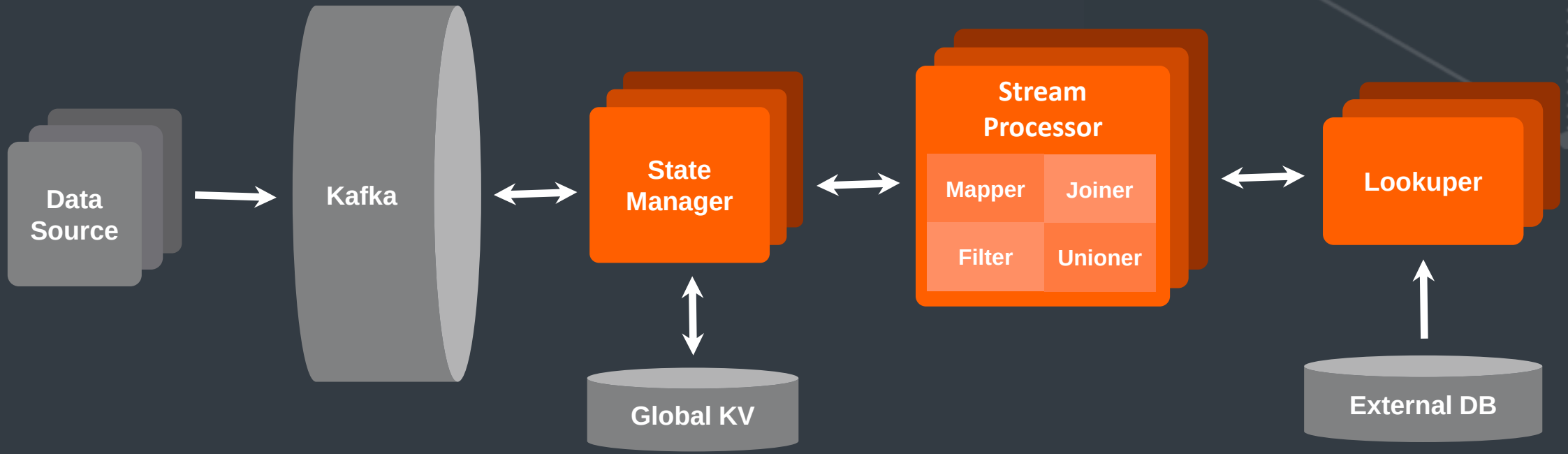
Streaming underlying

- Dojoyn
  - State Manager
  - Stream Reader
  - Stream Processor
    - Joiner
    - Mapper
    - Filter
    - Unioner
  - Lookuper
- Module act as an independent service
- Grpc between modules



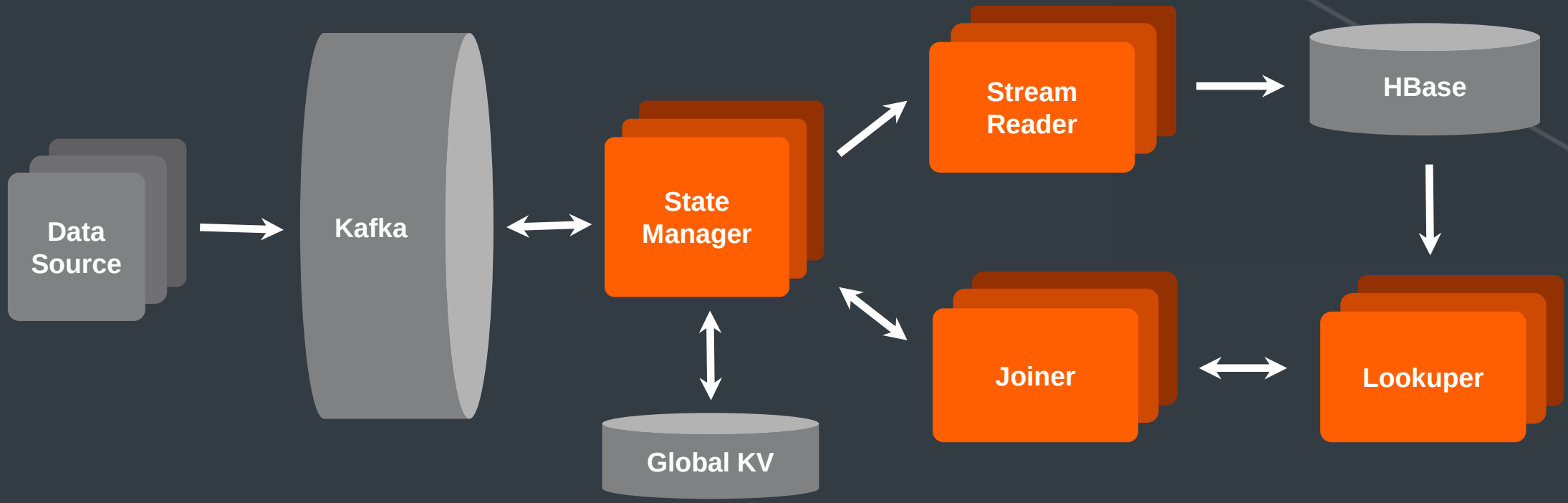
Single stream processing case

- Simple streaming without join
- Single stream join with static tables



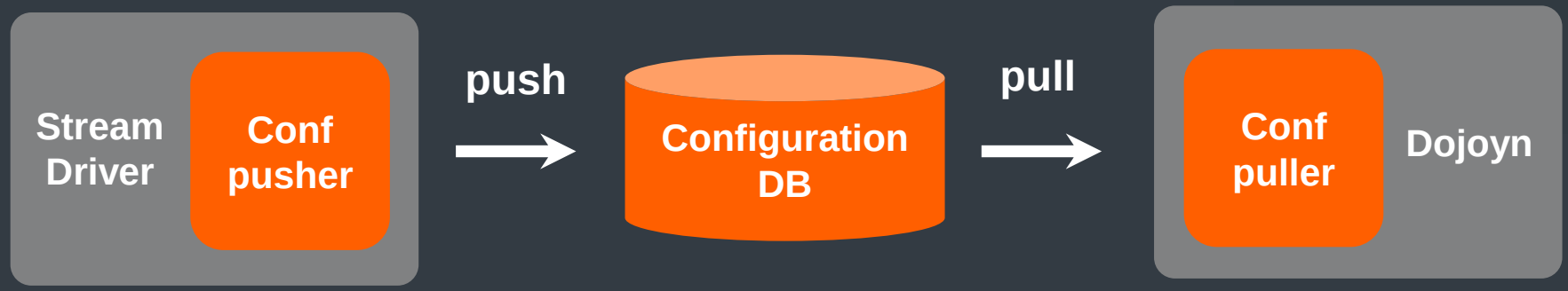
Multiple streams processing case

- Join between multiple streams
- Funnel events processing



Dynamic job creation/deletion

- Configure pusher & puller
- Data transferred with configure meta via rpc
- Data processed according with configure meta

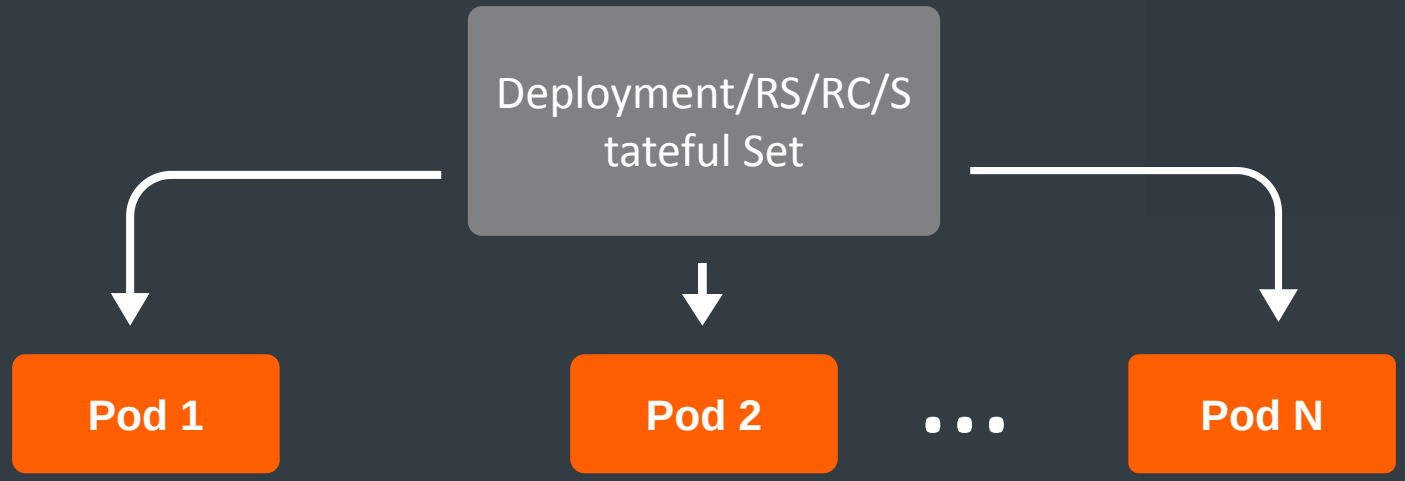


System design rules

- Scalability
- Load balance
- HA & Fault tolerance

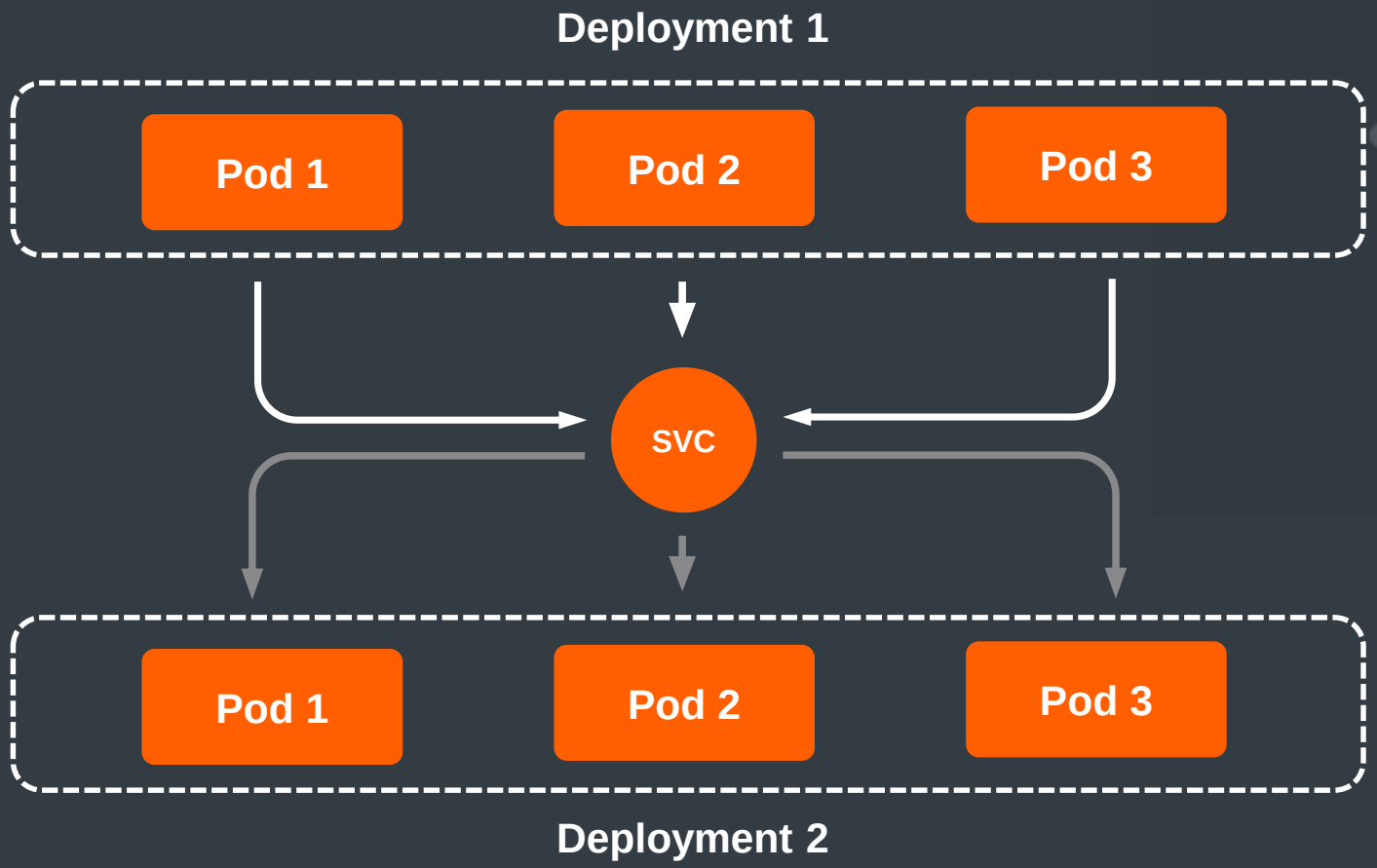
Scalability

- Supported by k8s
- One module, one Deployment
- Replication is freely set and can be change dynamically



Load balance

- k8s service



HA & Fault tolerance

- HA
  - k8s base software stack
  - Modules are stateless for streaming
  - Pod will automatically restart when crash
- Fault tolerance
  - Record state stored in global KV store
  - StateManager will retry for failed records

How the 4 questions are answered

- What **results are calculated?**
  - *Only element-wise for streaming, aggregation and composite done on druid side*
- Where **in event time are results calculated?**
  - *Originally calculated in event time*
- When **in processing time are results materialized?**
  - *No watermark and trigger needed, result will finally be correct*
- How **do refinements of results relate?**
  - *No watermark and trigger, so no refinements*

Case study*

- Case1

Select avg(a) from StreamA

Dojoyn

*Select * from StreamA*

Druid

Select avg(a) from TableA

- Case2

Select count(a) from StreamA join StreamB on a = b where Time < time1 and Time > time2

Dojoyn

*Select * from StreamA join StreamB on a=b*

Druid

Select count(a) from TableJ where Time < time1 and Time > time2

** HAP currently only support json query, here only use equivalent sql for better explain*

Streaming processing case*

```
SELECT * FROM demoapp da
JOIN ddt01 ddt1 ON da.IDTYP = ddt1.DOMVALUE_L AND ddt1.domname = 'IDTYP' AND ddt1.DDLANGUAGE = '1'
JOIN ddt01 ddt2 ON da.OPE_INS = ddt2.DOMVALUE_L AND ddt2.domname = 'OPERA' AND ddt2.DDLANGUAGE = '1'
JOIN ddt01 ddt3 ON da.OPE_DEVLOC = ddt3.DOMVALUE_L AND ddt3.domname = 'OPERA' AND ddt3.DDLANGUAGE = '1'
JOIN ddt01 ddt4 ON da.YPLACE = ddt4.DOMVALUE_L AND ddt4.domname = 'PLACE' AND ddt4.DDLANGUAGE = '1'
JOIN ddt01 ddt5 ON da.OPE_DEV = ddt5.DOMVALUE_L AND ddt5.domname = 'OPERA' AND ddt5.DDLANGUAGE = '1'
JOIN ddt01 ddt6 ON da.YCDIR = ddt6.DOMVALUE_L AND ddt6.domname = 'CDIR' AND ddt6.DDLANGUAGE = '1'
JOIN ddt01 ddt7 ON da.YFHKA = ddt7.DOMVALUE_L AND ddt7.domname = 'FHKA' AND ddt7.DDLANGUAGE = '1'
JOIN ddt01 ddt8 ON da.yndflg = ddt8.DOMVALUE_L AND ddt8.domname = 'NDFLG' AND ddt8.DDLANGUAGE = '1'
JOIN ddt01 ddt9 ON da.ZOPTYP = ddt9.DOMVALUE_L AND ddt9.domname = 'OPTYP' AND ddt9.DDLANGUAGE = '1'
JOIN ddt01 ddt10 ON da.METYP = ddt10.DOMVALUE_L AND ddt10.domname = 'METYP' AND ddt10.DDLANGUAGE = '1'
JOIN ddt01 ddt11 ON da.YCMET = ddt11.DOMVALUE_L AND ddt11.domname = 'CMET' AND ddt11.DDLANGUAGE = '1'
JOIN ddt01 ddt12 ON da.OPE_CO = ddt12.DOMVALUE_L AND ddt12.domname = 'OPERA' AND ddt12.DDLANGUAGE = '1'
JOIN ddt01 ddt13 ON da.OPE_MR = ddt13.DOMVALUE_L AND ddt13.domname = 'OPERA' AND ddt13.DDLANGUAGE = '1'
JOIN ddt01 ddt14 ON da.OPE_COM = ddt14.DOMVALUE_L AND ddt14.domname = 'OPERA' AND ddt14.DDLANGUAGE = '1'
JOIN ddt01 ddt15 ON da.FSTACT = ddt15.DOMVALUE_L AND ddt15.domname = 'FSTACT' AND ddt15.DDLANGUAGE = '1'
JOIN ddt01 ddt16 ON da.OPE_EXCH = ddt16.DOMVALUE_L AND ddt16.domname = 'OPERA' AND ddt16.DDLANGUAGE = '1'
JOIN ddt01 ddt17 ON da.YFHKA_O = ddt17.DOMVALUE_L AND ddt17.domname = 'YHKA' AND ddt17.DDLANGUAGE = '1'
JOIN ddt01 ddt18 ON da.YCDIR_O = ddt18.DOMVALUE_L AND ddt18.domname = 'CDIR' AND ddt18.DDLANGUAGE = '1'
JOIN TBT03 hyxt ON da.ISTYPE = hyxt.ISTYPE AND hyxt.LANGU = '1' AND hyxt.CLIENT = '800'
JOIN TBB03 hy ON da.ISTYPE = hy.ISTYPE AND hy.IND_SECTOR = da.BRANCHE AND hy.SPRAS = '1' AND hy.CLIENT = '800'
JOIN qmel ajd ON da.idnum = ajd.qmnum AND ajd.mandt = '800'
WHERE da.MANDT = '800' AND da.IDTYP = 'T' AND da.SERNO = '1' AND da.SDENO = '1'
```

** HAP currently only support json query, here only use equivalent sql for better explain*

Lower latency & higher throughput

- 6 million records join with 50 million records for 20+ times
- Hive complete with 72+ minutes
- Dojoyn done within 120s
- 120s VS 72+ minutes, 36X faster, much lower latency

Conclusion

- User friendly
- Good performance
- Flexible query window
- Multiple streams join support
- Multiple jobs support
- Advanced OLAP query support

Future work

- SQL support
- More streaming engine support (session window)
- Better query optimization
- Elastic scaling

Q&A

Address:
Room 1002, Bldg C, No. 421, Zhengli Rd, Yangpu District, Shanghai, China

Phone: +86 15900532616
Direct line: +86 21 33620185
Fax: +86 21 33620185
Email: liyezhang556520@gmail.com
Web: http://he2.io

Thanks
谢谢！