traveloka

The Era of Big Data

Andi N. Dirgantara

Speaker Profile

traveloka

- I’m Andi Nugroho Dirgantara
- More than 6 years as software engineer
- Last 4 years focused at data engineering (big data)
- Lead Data Engineer at Traveloka
- Lead Facebook Developer Circles Malang
- Co-founder The Bros Coffee and Coworking Space (@thebros_co)
- Co-founder Cahayu Aesthetic and Slimming Center (@cahayu.clinic)
- Working remotely from Malang

What is the problem?

The way we
store, **process**, and **consume** data
has been changed

Imagine this problem

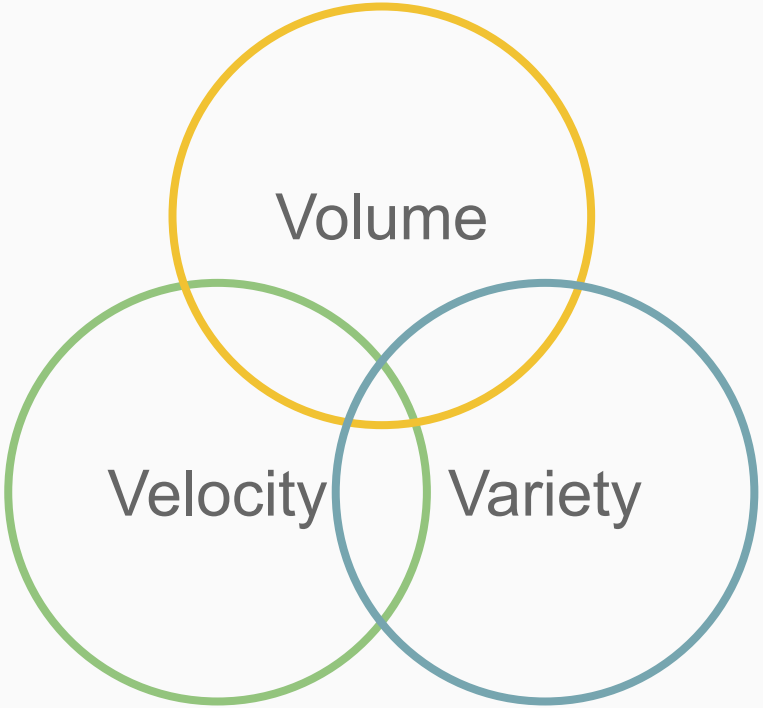
traveloka

We have MySQL installed on some cloud instance, used by our application/ service

Everything went well until…

- We faced **50,000 rows per seconds** (18 millions rows per hour)
- Storage consume more than **100GB each days**
- Single **query** can takes more than **5 hours**

3V of Big Data



Some references add **variety** and **value** so it becomes 5V

What is the solution?

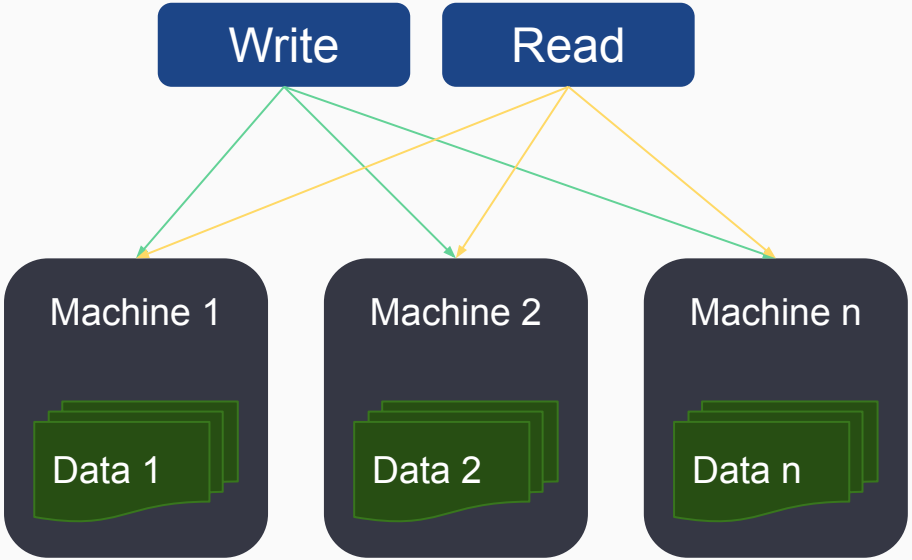
Use **distributed system**

to leverage **horizontal scalability**

Distributed system is the way
Partition is the key

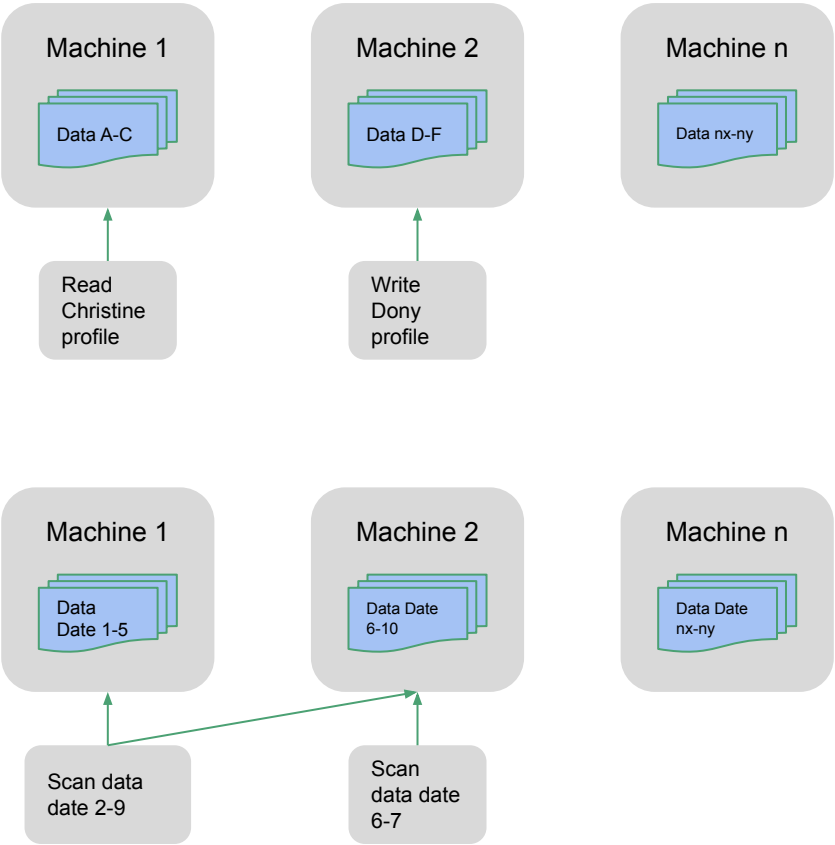
Throughput
Storage
Query
} problems, is because single machine do it all together

So let's break it down to multiple machine instead.



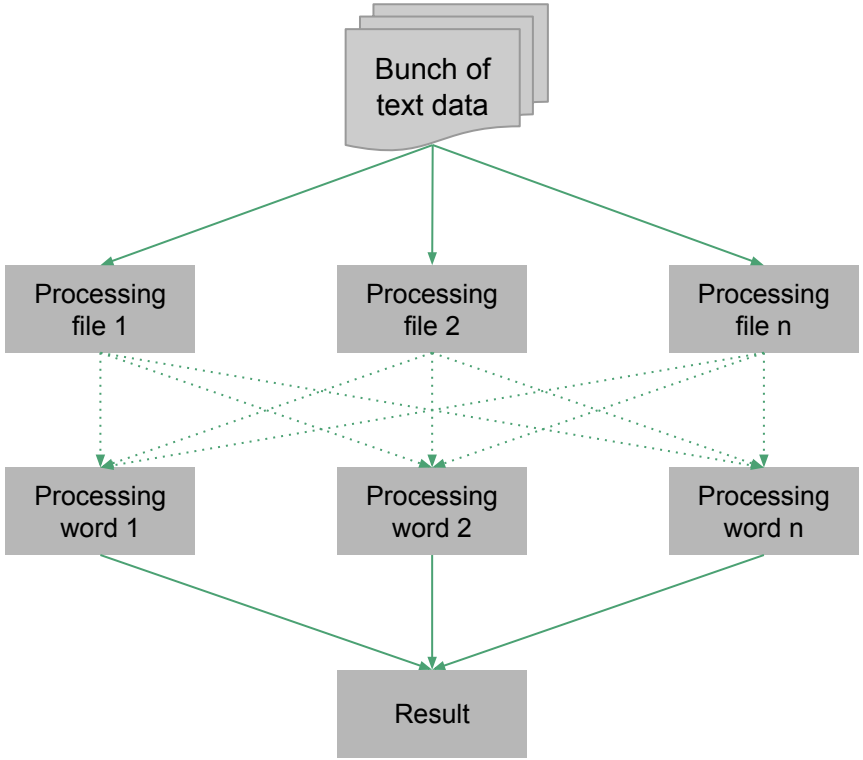
How

Storing in distributed system
Example storing user profile



How

Processing in distributed system
Example doing word count



Our problems are solved!

Throughput
Storage
Query

} All of those problems were solved

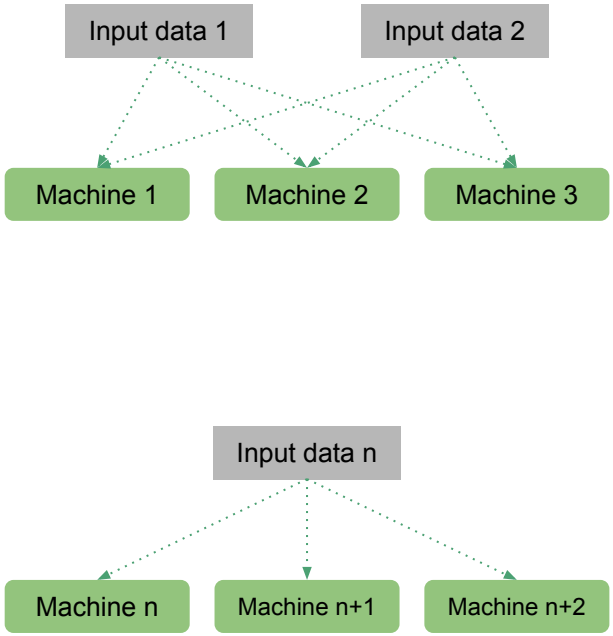
but another problem then raised...

What if?

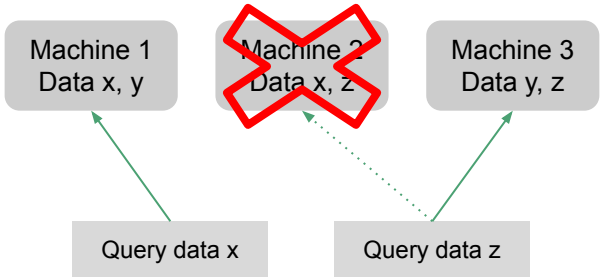
One machine was down

We needs high availability

Replication will solve it
Example replication factor 3

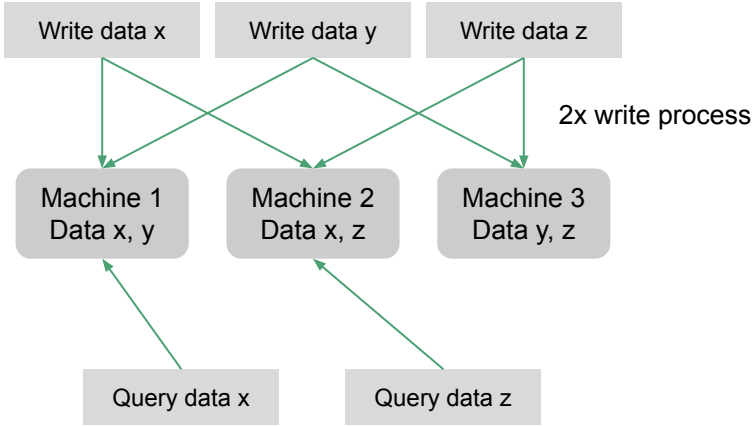


High availability solves hardware failure



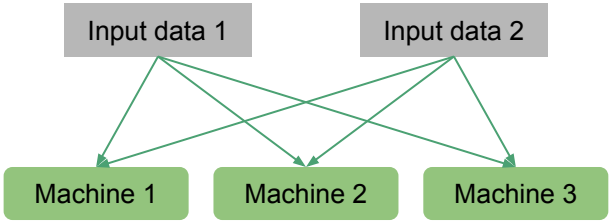
another problem raised again...

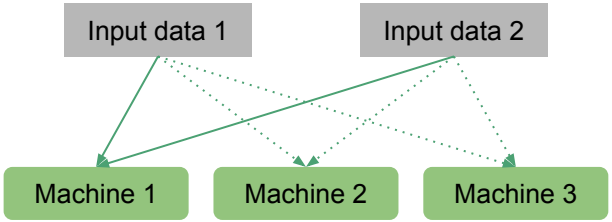
Replication factor slowing down I/O process



We needs
consistency control

We can choose success sign either all, quorum, or only one replication machines said succeed





Now we have

- Partition system
- Replication factor
- Consistency control

Who’s responsible to manage it?

Data Team Organization

Data Team Specialization

traveloka

Data Analyst

Common tasks

- Data visualization
- Data processing
- Applying statistical analysis
- Provide report to other division (marketing, etc.)

Common technology stacks:

- Business Intelligence Tools (Power BI, Pentaho, etc.)
- Data visualization (Domo, Periscope, etc.)
- Spreadsheet (advance usage)

Data Engineer

Common tasks:

- Setup data warehousing infrastructure
- Performing Extract Transform and Load (ETL)
- Creating internal tooling for internal data team

Common technology stacks:

- Hadoop family (HDFS, HBase, Zookeeper, etc.)
- Distributed processing framework (Spark, Beam)
- Distributed storage (Cassandra, Sharded Mongo, etc.)

Data Scientist

Common tasks:

- Performing machine learning and artificial intelligence stuff
- Predictive data modelling
- Advance statistical analysis

Common technology stacks:

- Machine learning framework (TensorFlow, Pytorch, etc.)
- Any Python library for data scientist (Numpy, Panda, etc.)
- Distributed ML framework (SparkML)

Any example above based on experience from some start ups, there's no formal guidelines for data team specialization, so the implementation inside industry may vary.

Where should I start?

Both reading a book and practicing is important

Literacy

- Read articles, books, and any reference for data engineering.
- The keywords usually “big data”, “hadoop”, “spark”, “cassandra”, “elastic search”, and any other big data tech. stacks.
- Watching videos also works, a lot of big data reference on YouTube and any other video courses.

Experience

- Knowing the concept without implementation is dull.
- Some practical approaches in industry usually leveraging higher level abstraction, so we need to be used to it if we want to make an impact in industry.

Data Engineering References

traveloka

- http://highscalability.com/
- Following big data vendor **social media** like Cloudera, Hortonworks, etc.
- **Medium, Quora**, or any user generated content website.
- **Documentation** on every technology stacks.
- **Open source** project discussion.
- **Conferences** video.
- Many more...

Common Pitfalls

traveloka

- **The company is actually doesn't need any big data solution**
  Don't follow the hype, big data is not a silver bullet, even most cases doesn't well fit with big data solution.
- **The data is not big enough**
  Similar to the first point, when the data is not big already, use existing solution is preferred.
- **Company allocation for data infrastructure cost is below the requirement**
  For certain point when we need to implement big data, make sure the allocation budget is able to cover its cost.
- **Don't have good mentor**
  Since the technology or concept itself is new, there are not so much senior data engineer out there.

Start from Small

traveloka

- **Start from small** then iterate often, is always a good approach.
- Grow your company **data infrastructure** as well as your **team capability**.
- **Never worried** about **data migration**.
  It's one of common data engineering tasks. Don't take that role if you're afraid of it.

Where to go next?

Do higher impact action

traveloka

- **Contribute in open source library/ framework**
  It will help us always up to date with the new technology as well as improving our technical capability.
- **Write some books or articles in any medium**
  Helps us remembering things and help other people learn too.
- **Share the knowledge**
  Active in community, actively speaking at tech conferences/ meetups, help us expand our networking, multiplied the impact.

Join

Developer Circles
from facebook

http://bit.ly/DevCMalang

We are hiring

Traveloka Data Engineer Opportunity

traveloka

Data Integration Engineer Engineering • Full-time	Indonesia	Details
Data Site Infrastructure Engineer Engineering • Full-time	Indonesia	Details
Data Engineer Engineering • Full-time	Indonesia	Details
Data Engineer Engineering • Full-time	Singapore	Details
Data Integration Engineer Engineering • Full-time	Indonesia	Details

Contact me if you're interested to join us

traveloka

Thank you and see you again!

Let’s keep in touch!

fb.me/andi.n.dirgantara

andi_dirgantara

hellowin