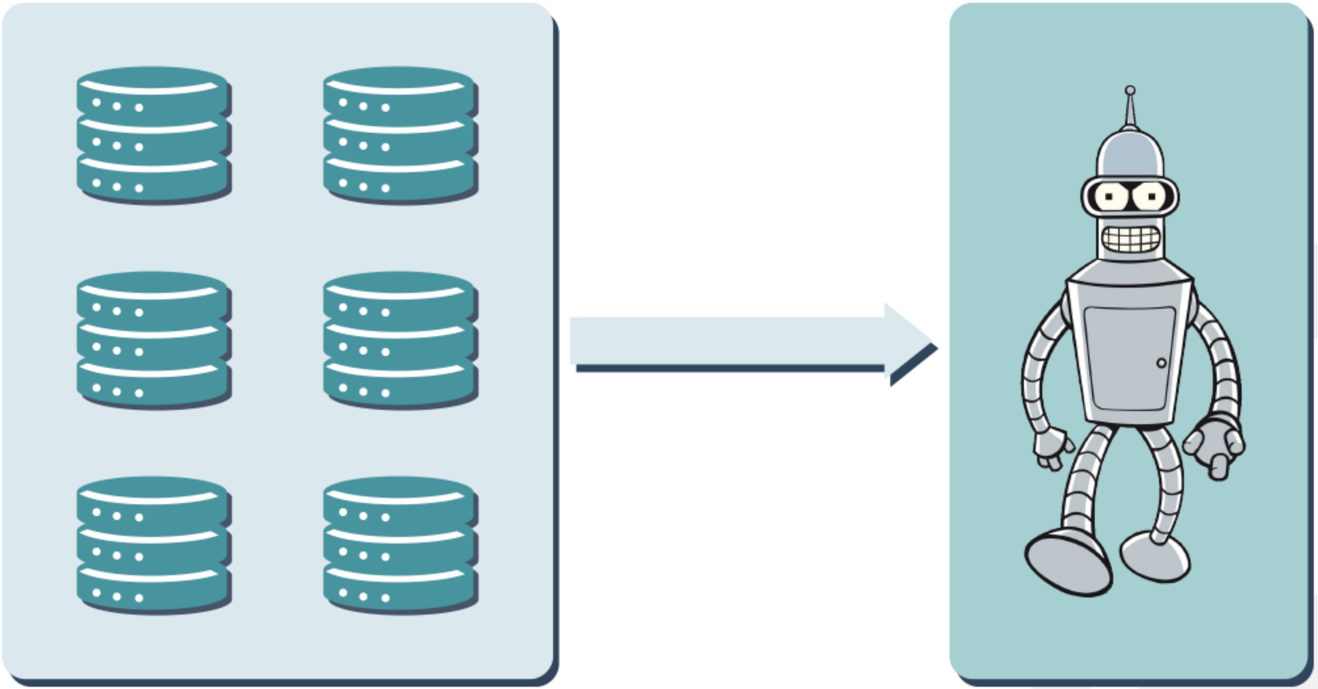
Distributed Machine Learning with Zero ETL

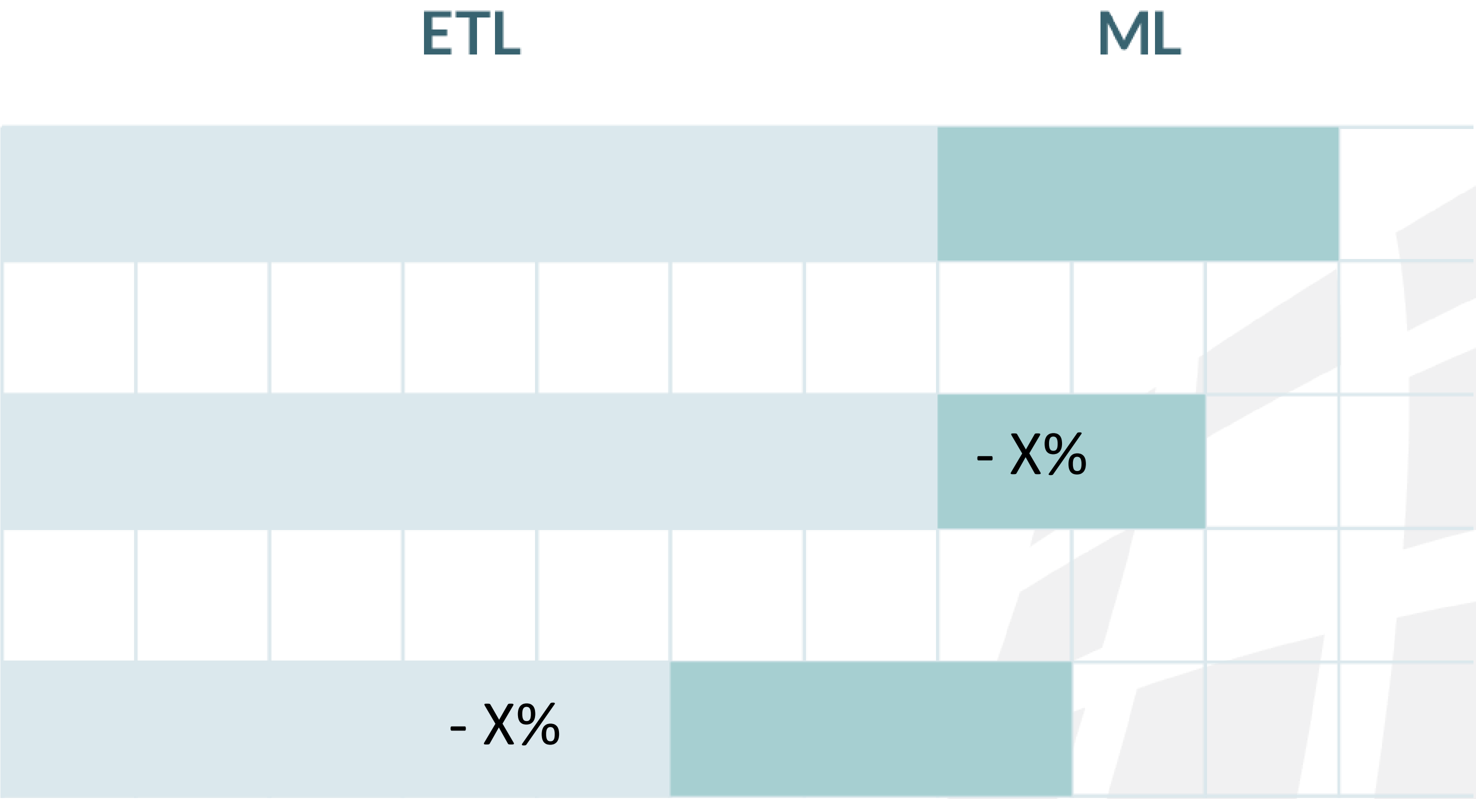
Yury Babak

Head of development, GridGain

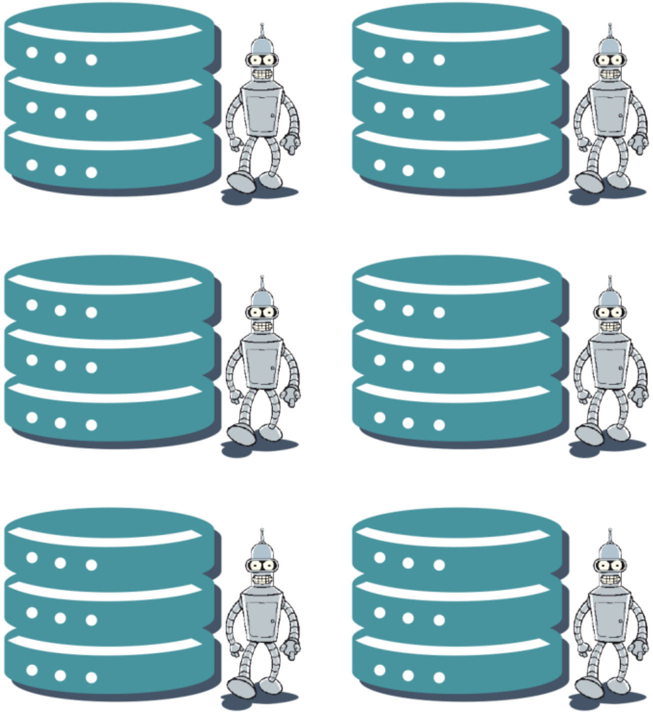
Long ETL



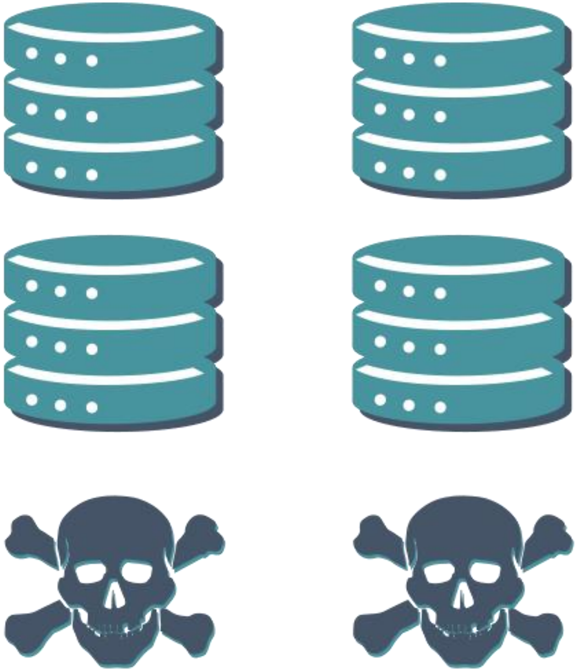
Long ETL



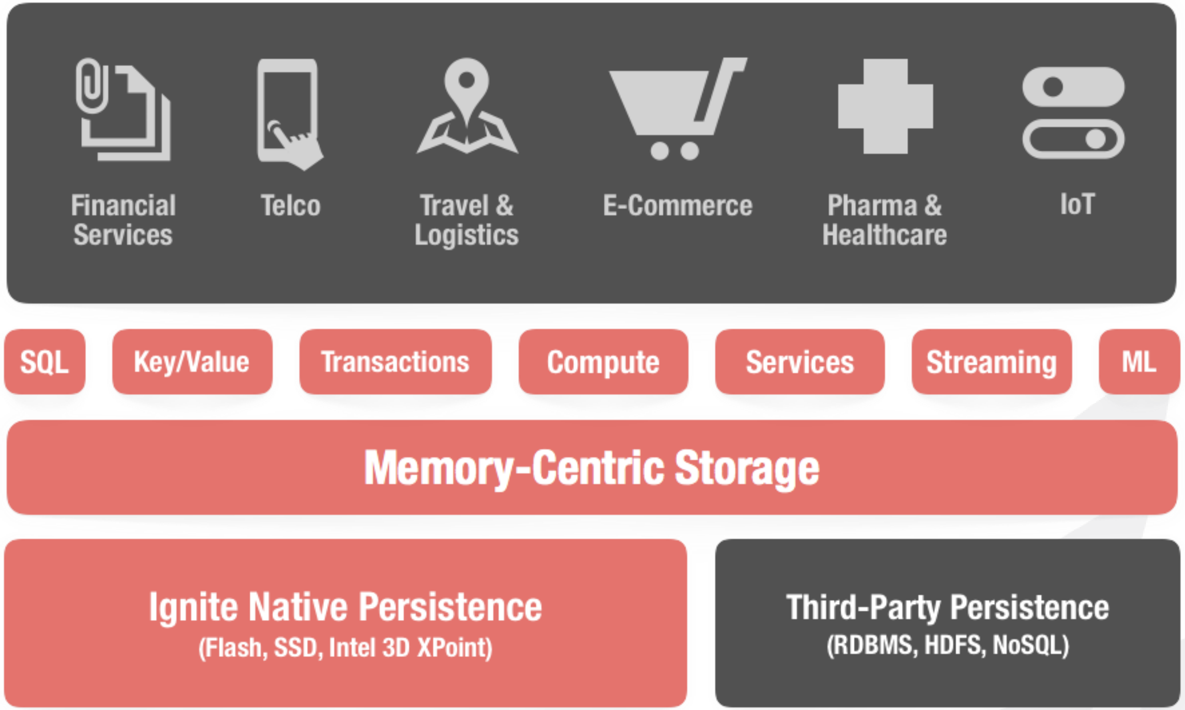
Distributed Training



Node Crash



Apache Ignite



Financial Services

Telco

Travel & Logistics

E-Commerce

Pharma & Healthcare

IoT

SQL

Key/Value

Transactions

Compute

Services

Streaming

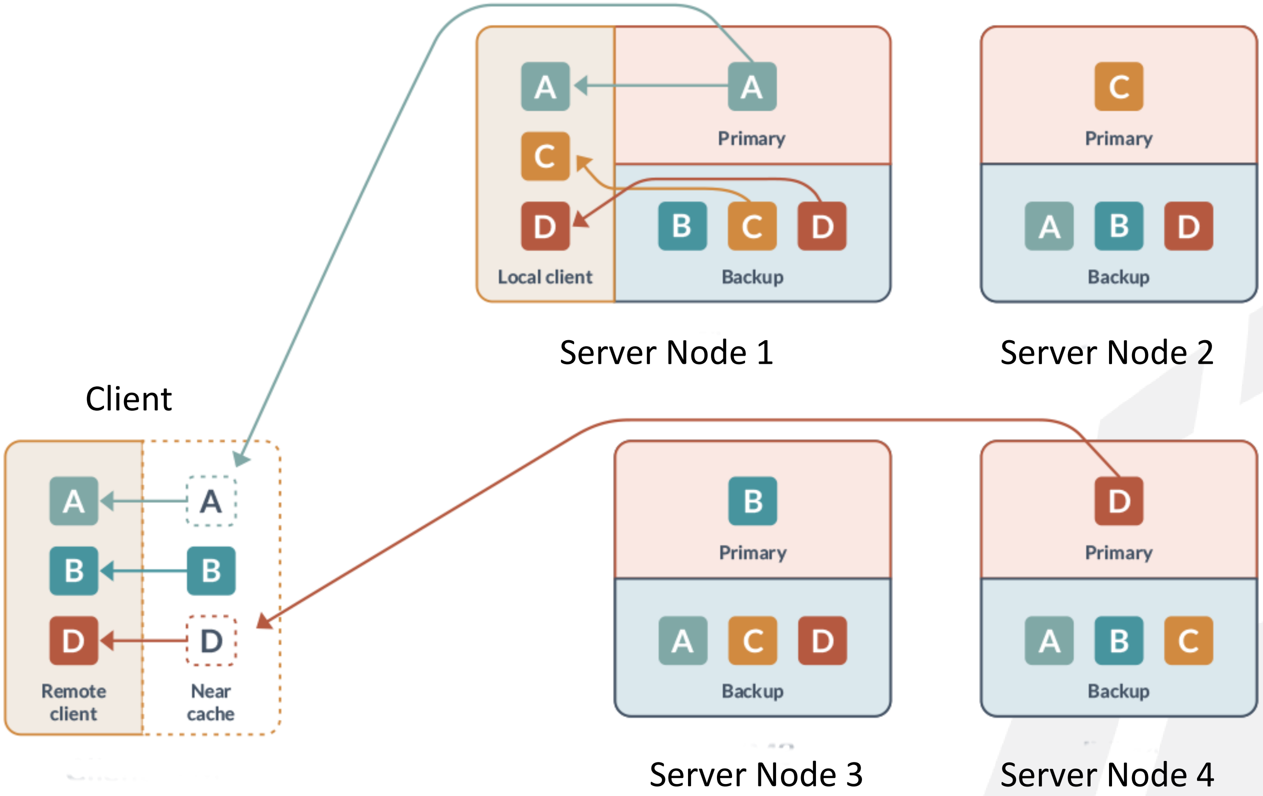
ML

Memory-Centric Storage

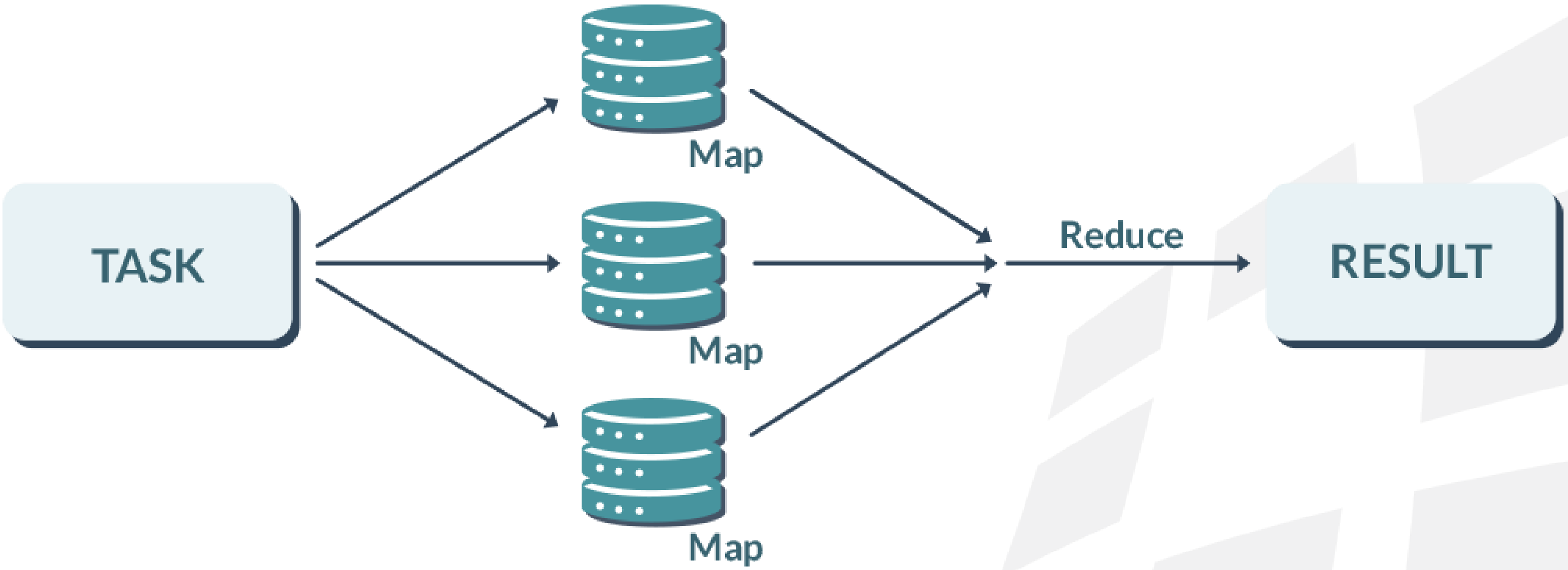
Ignite Native Persistence
(Flash, SSD, Intel 3D XPoint)

Third-Party Persistence
(RDBMS, HDFS, NoSQL)

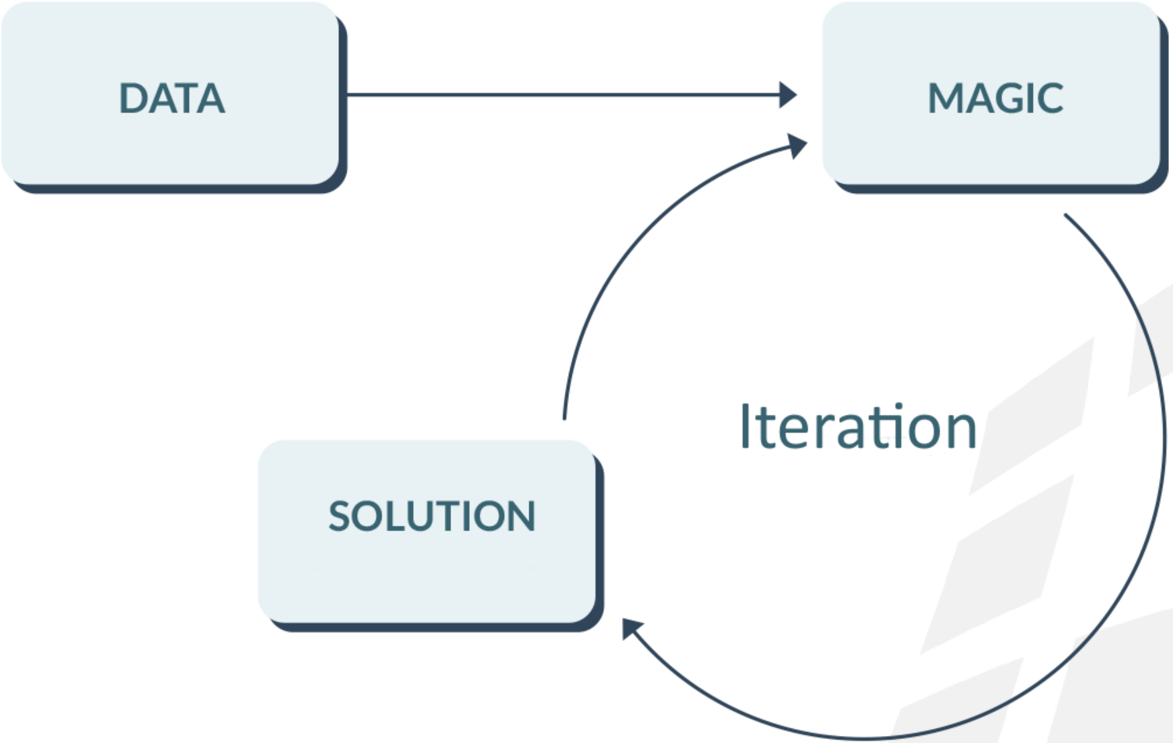
Apache Ignite: Replicated Caches



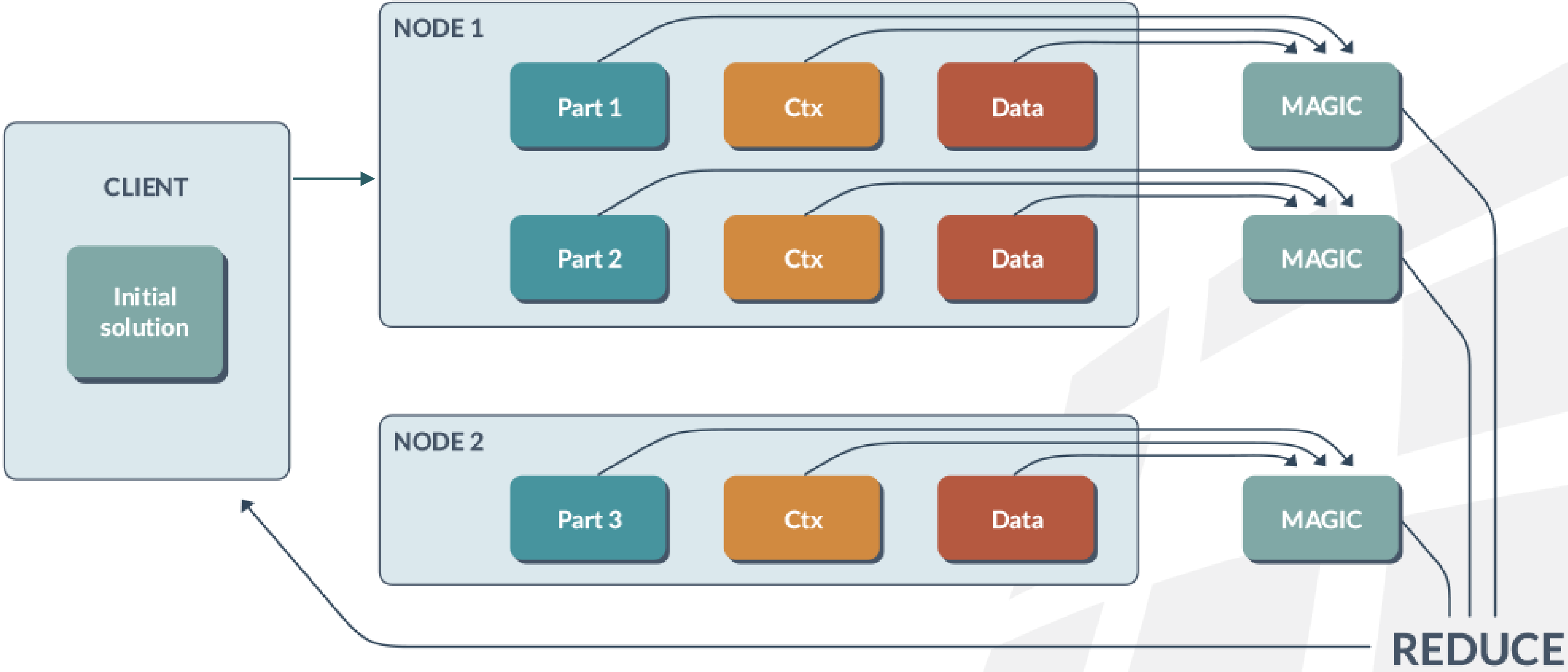
Map Reduce



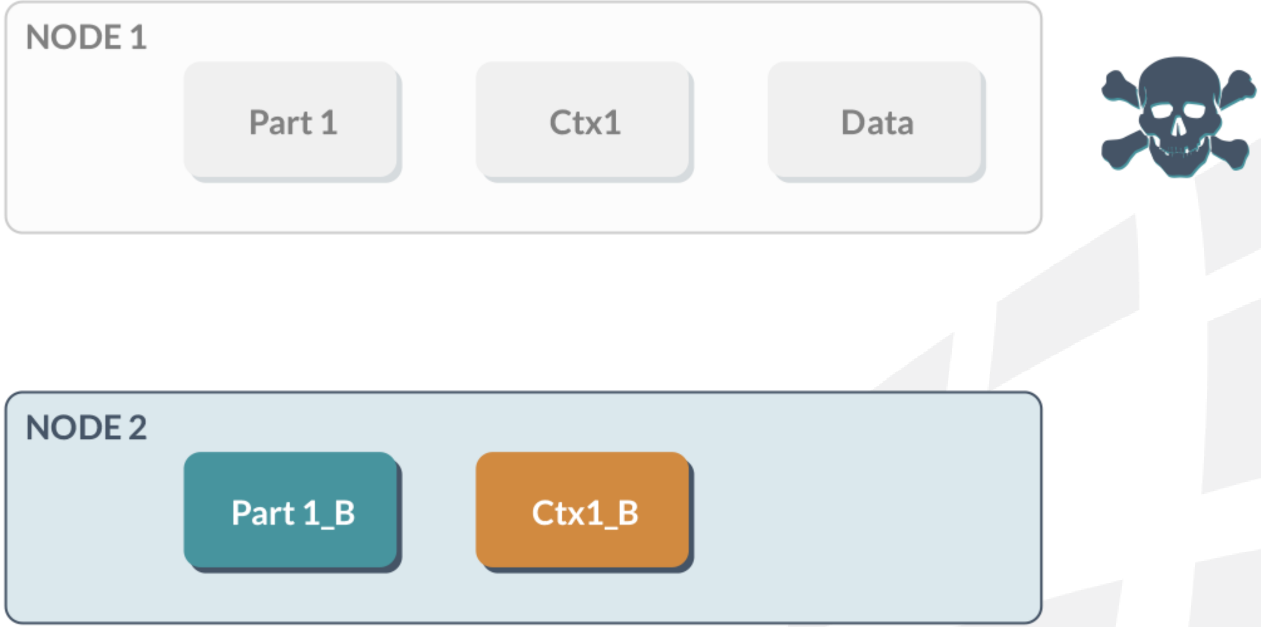
Iterative Optimization Algorithm



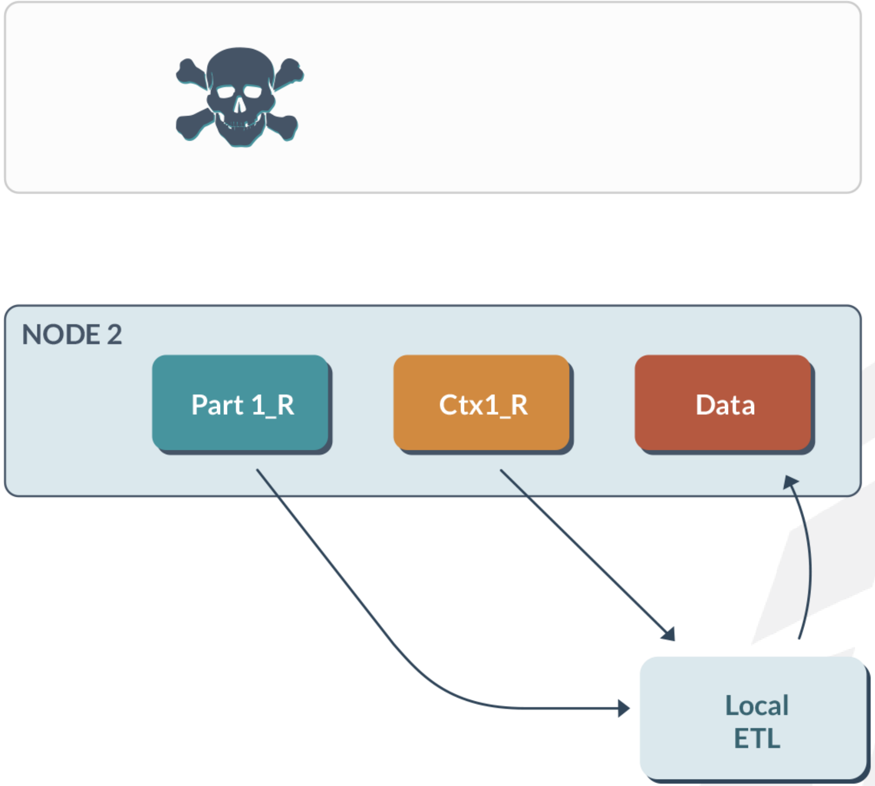
Partition Based Data Set



Restoration of partitions after a failure



Recovering calculations after failure



OLS sample

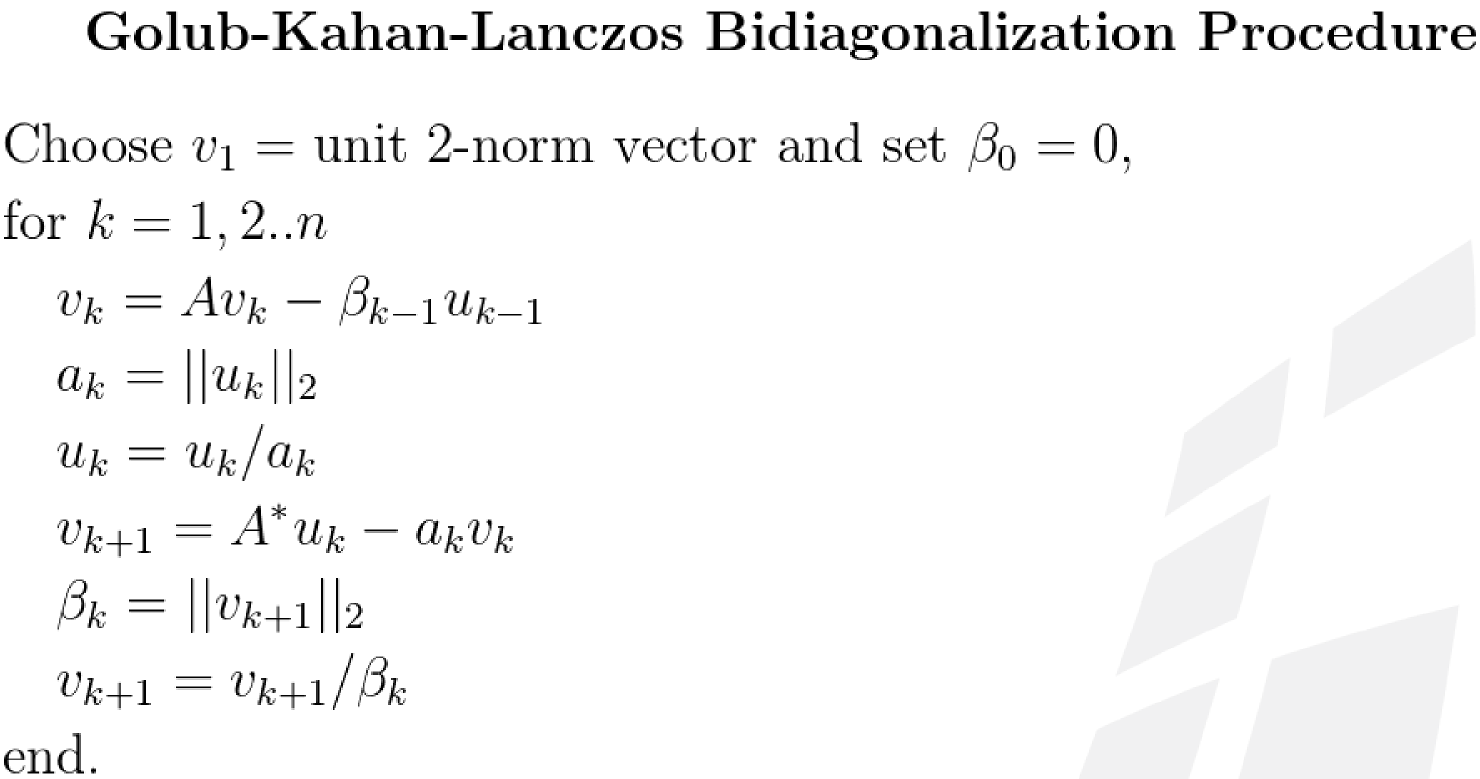
Loss function

$$(f(x_1)-y_1)^2+(f(x_2)-y_2)^2+\cdots+(f(x_n)-y_n)^2$$

Gradient of loss function

$$\underbrace{2(f(x_1)-y_1)*f'(x_1)}_{\text{Node 1}}+\underbrace{2(f(x_2)-y_2)*f'(x_2)}_{\text{Node 2}}+\cdots+\underbrace{2(f(x_n)-y_n)*f'(x_n)}_{\text{Node M}}$$

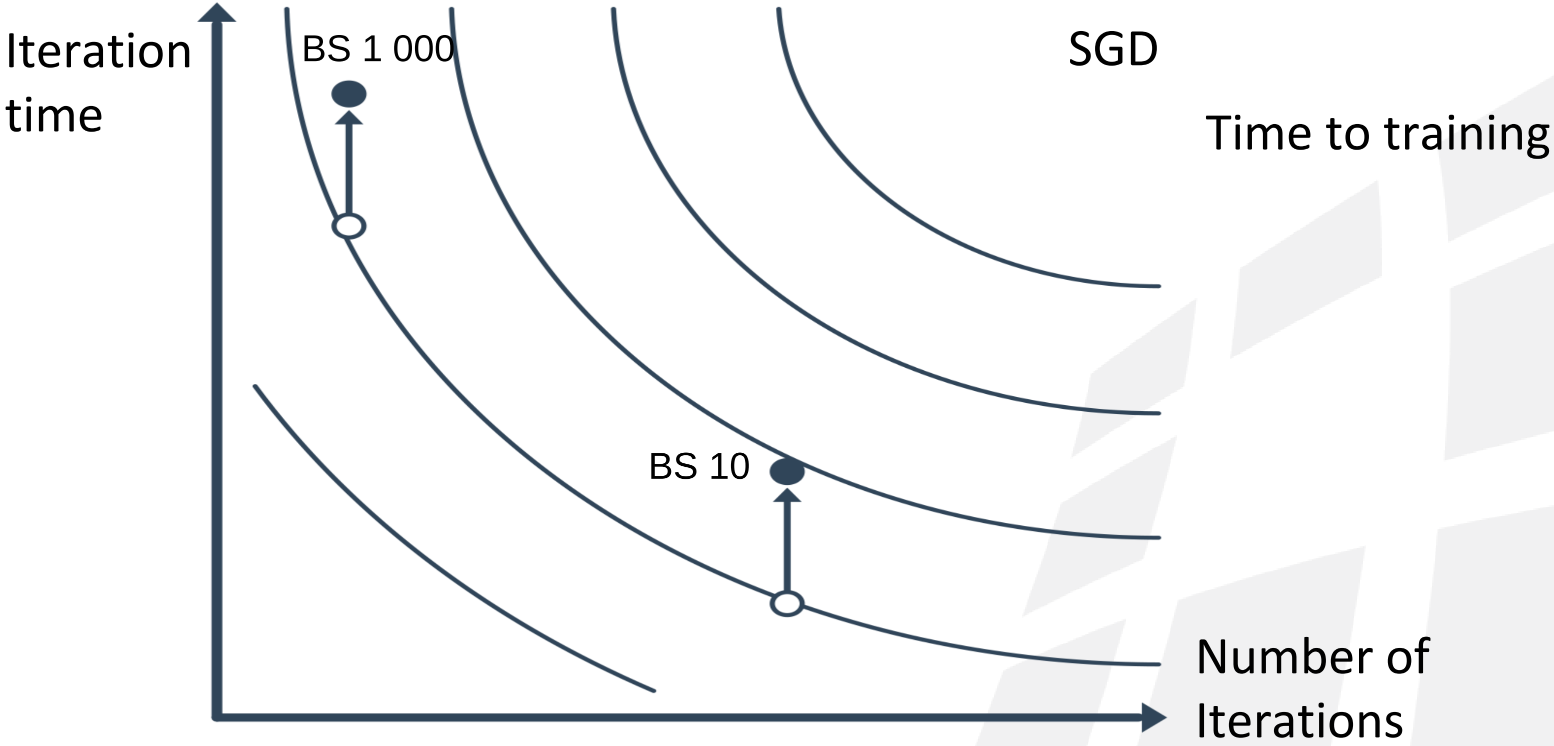
Sample 2 LSQR



Golub-Kahan-Lanczos Bidiagonalization Procedure

```
Choose $v_1$ = unit 2-norm vector and set $\beta_0 = 0$,
for $k = 1, 2..n$
    $v_k = Av_k - \beta_{k-1}u_{k-1}$
    $a_k = ||u_k||_2$
    $u_k = u_k/a_k$
    $v_{k+1} = A^*u_k - a_k v_k$
    $\beta_k = ||v_{k+1}||_2$
    $v_{k+1} = v_{k+1}/\beta_k$
end.
```

Limitations of Applicability



Want to learn more?

https://ignite.apache.org

https://apacheignite.readme.io/docs

https://github.com/apache/ignite

ybabak@gridgain.com