

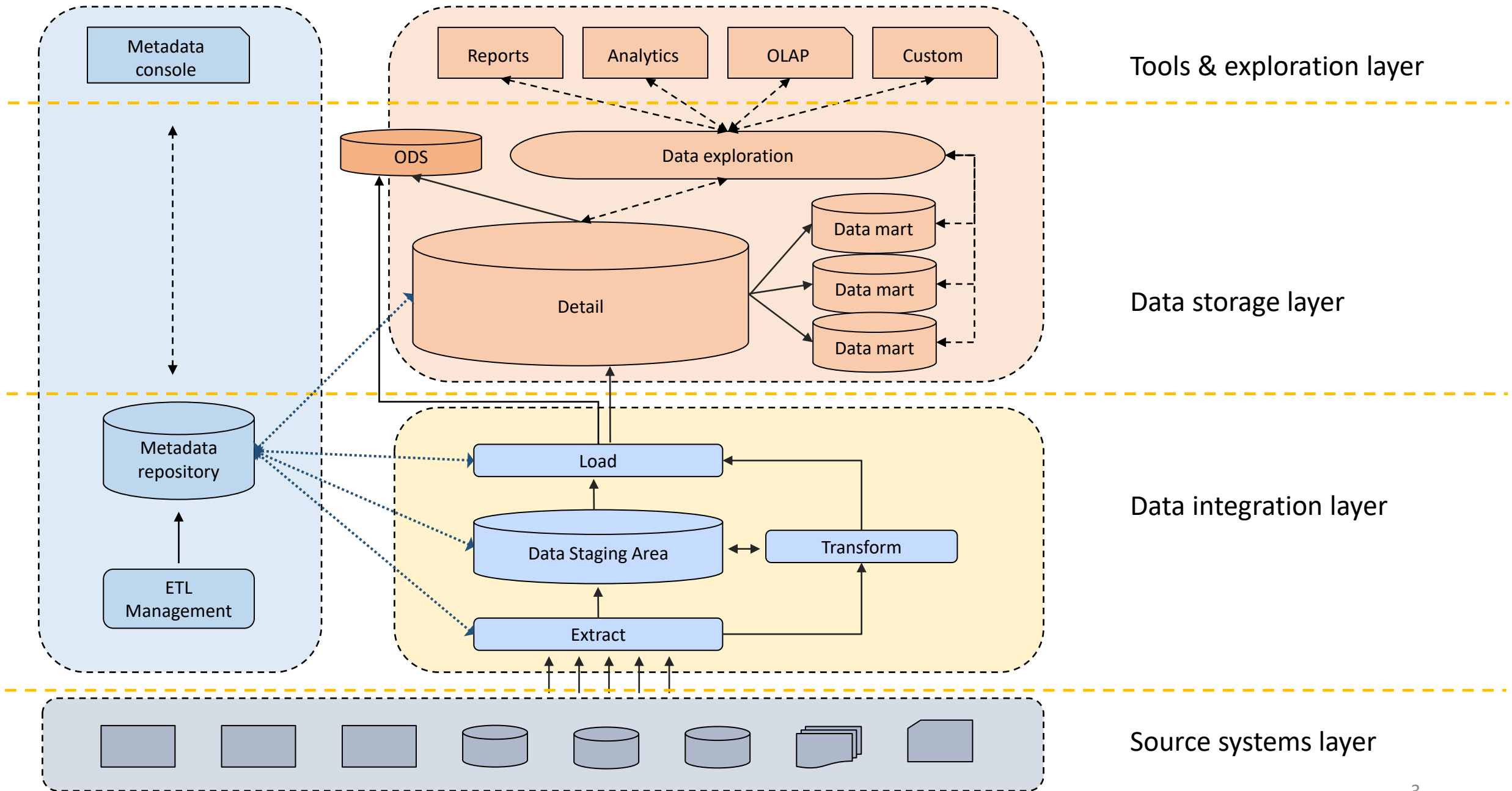
Hadoop i Spark

Mariusz Rafało

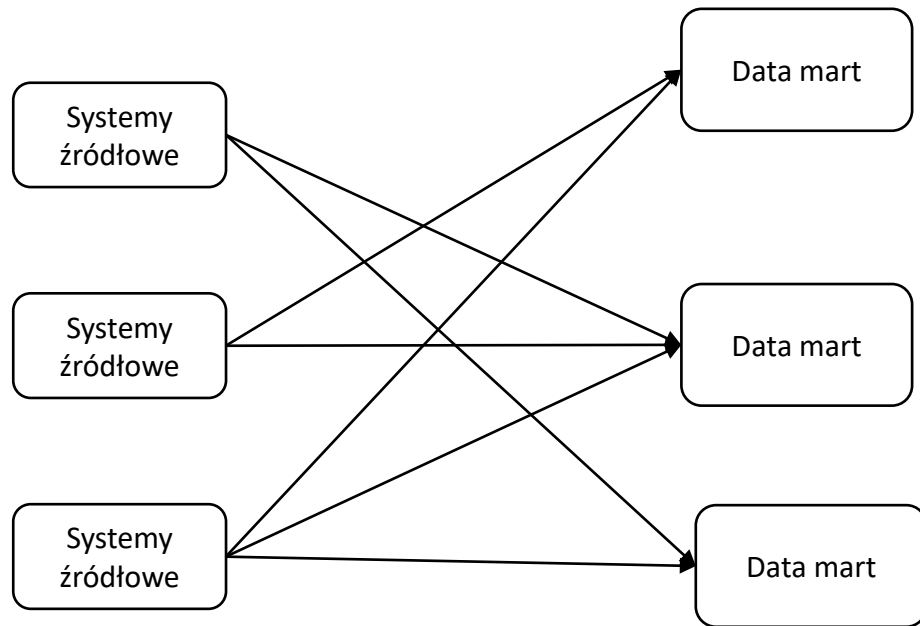
mrafalo@sgh.waw.pl

<http://mariuszrafalo.pl>

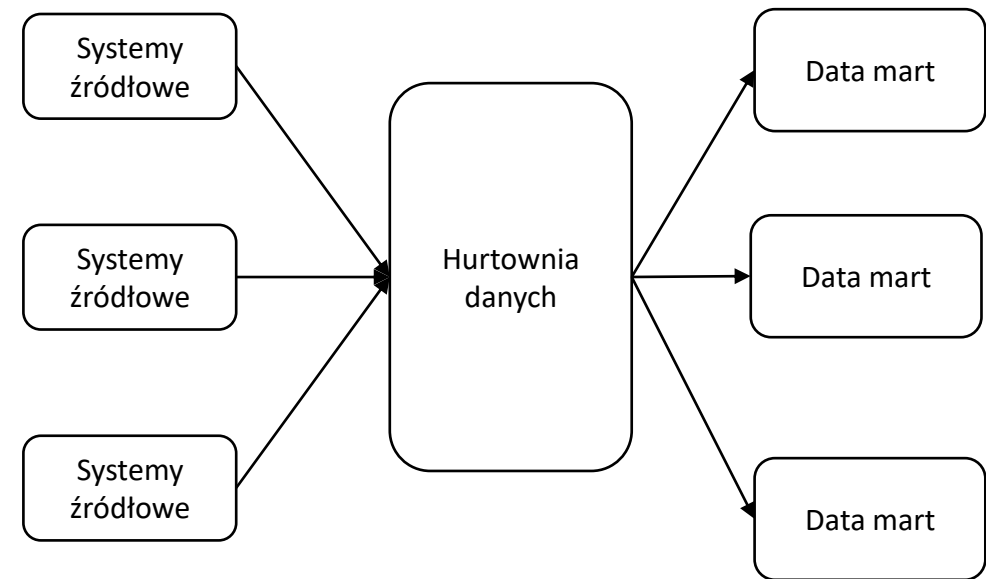
ARCHITEKTURY TRADYCYJNYCH HURTOWNI DANYCH



Warianty architektury HD (1)

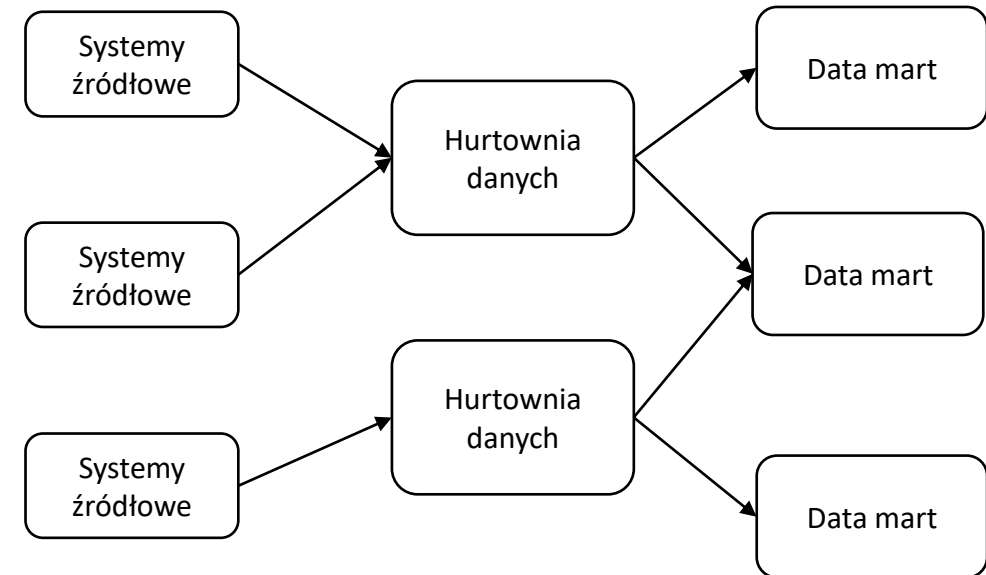
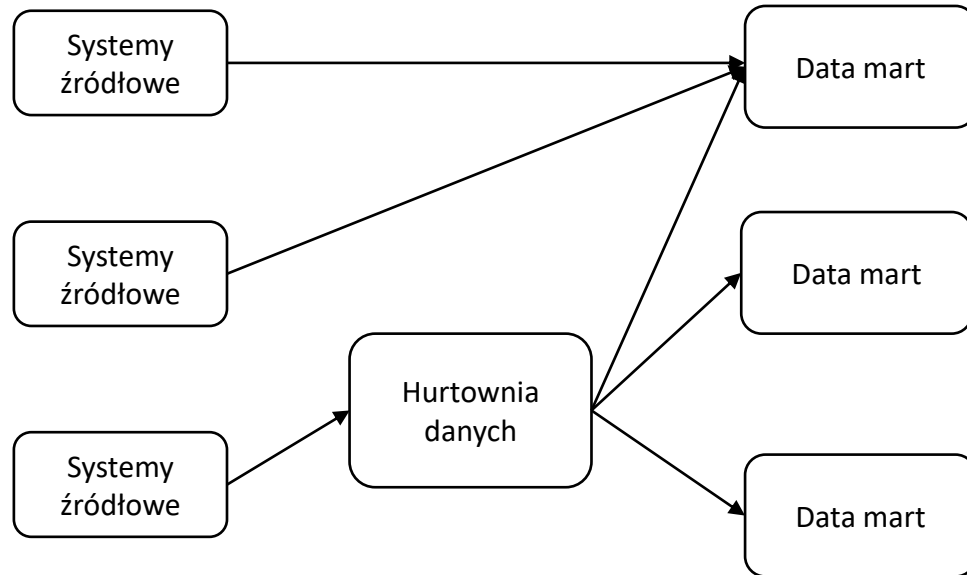


"...the data warehouse is nothing more than the union of all the data marts,"
R.Kimball

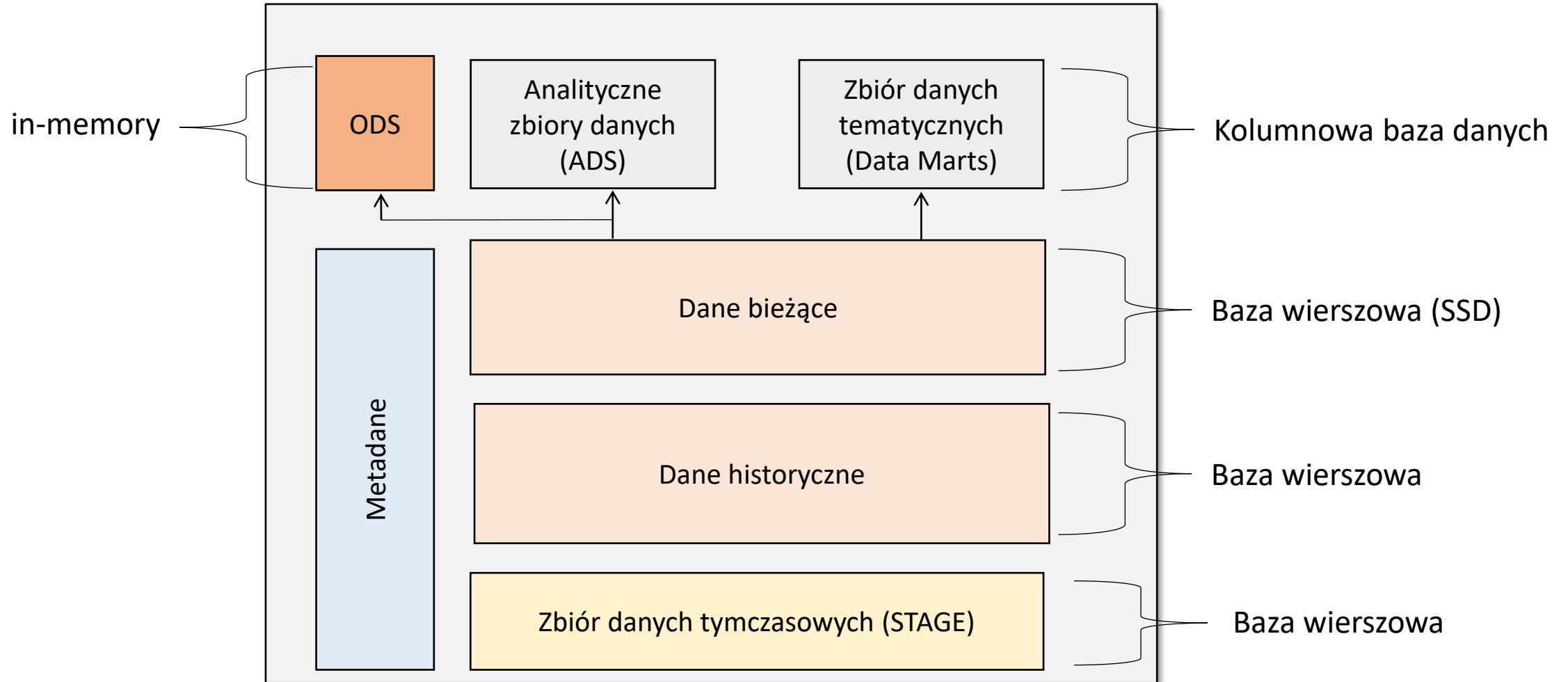


"You can catch all the minnows in the ocean and stack them together and they still do not make a whale,"
W.Inmon

Warianty architektury HD (2)



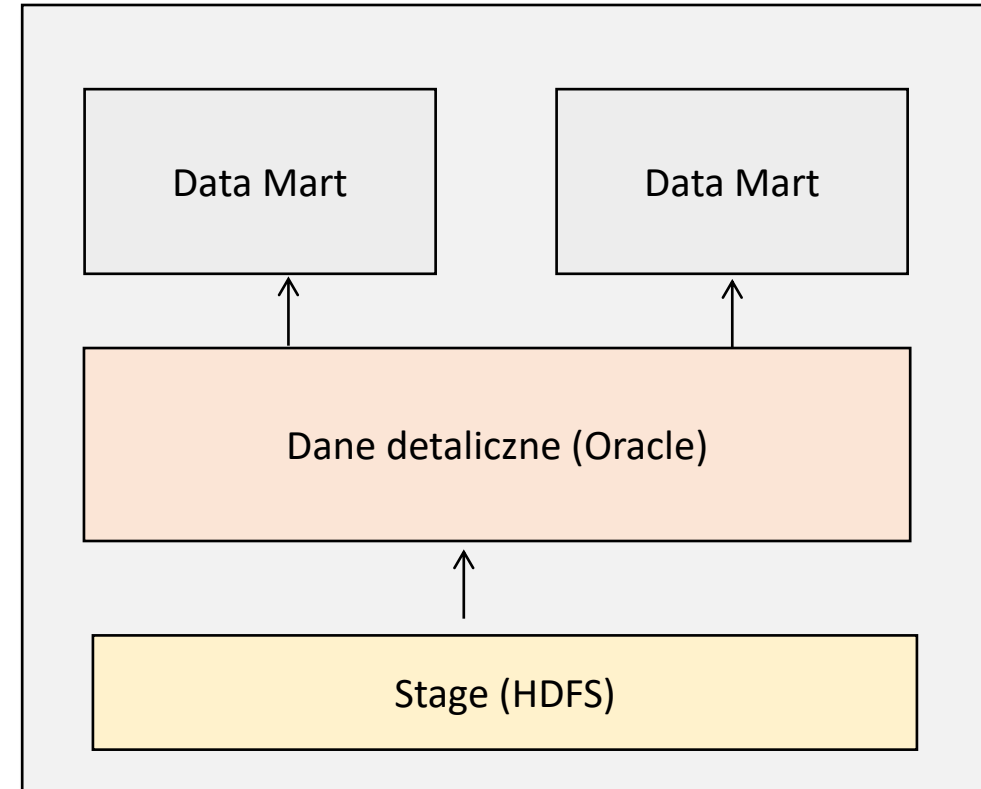
Technologie baz danych



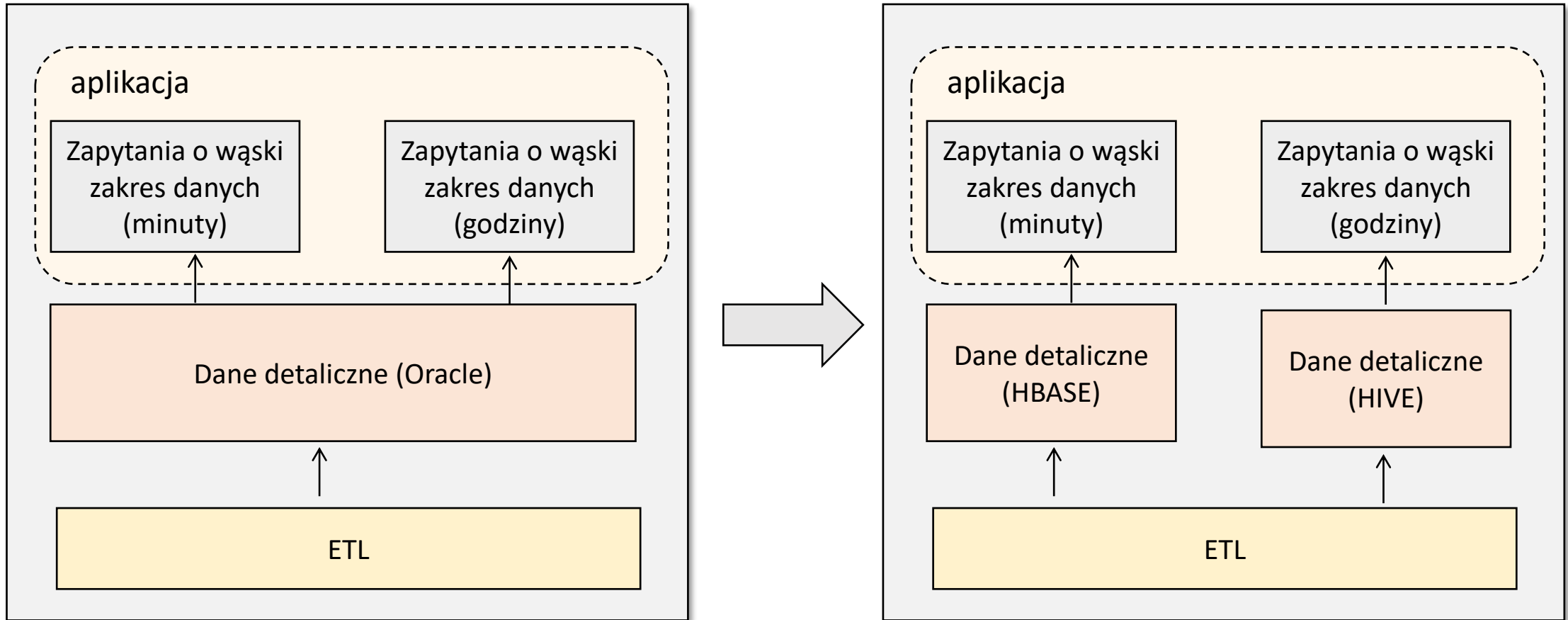
SPARK i HADOOP W EKOSYSTEMIE HURTOWNI DANYCH

Przykład: big data jako stage w HD

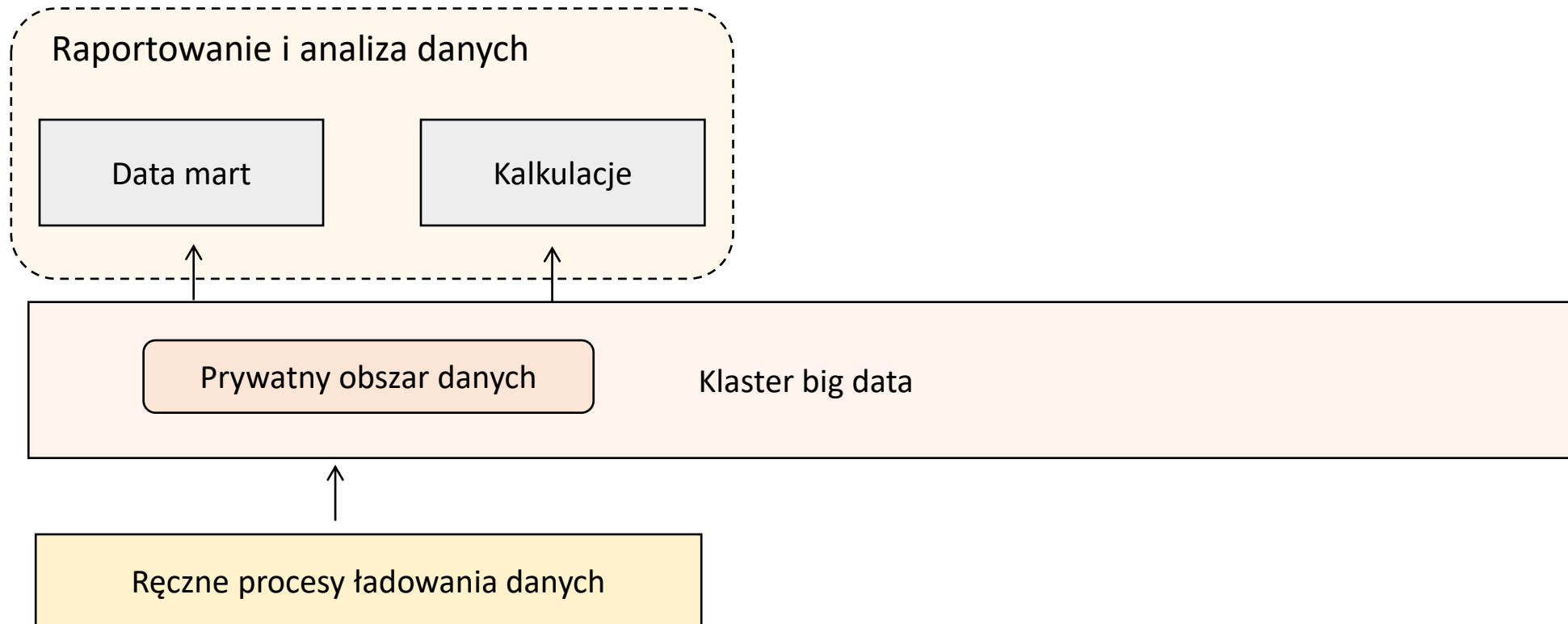
- Część procesu ETL oraz warstwa stage zostały zastąpione przetwarzaniem na platformie hadoop
- Na HDFS trafiają pliki, które poddawane są transformacji i ładowaniu do warstwy danych detalicznych
- HDFS utrzymuje także dane wymagane na potrzeby przyszłych przeliczeń danych, lub wzbogacania danych



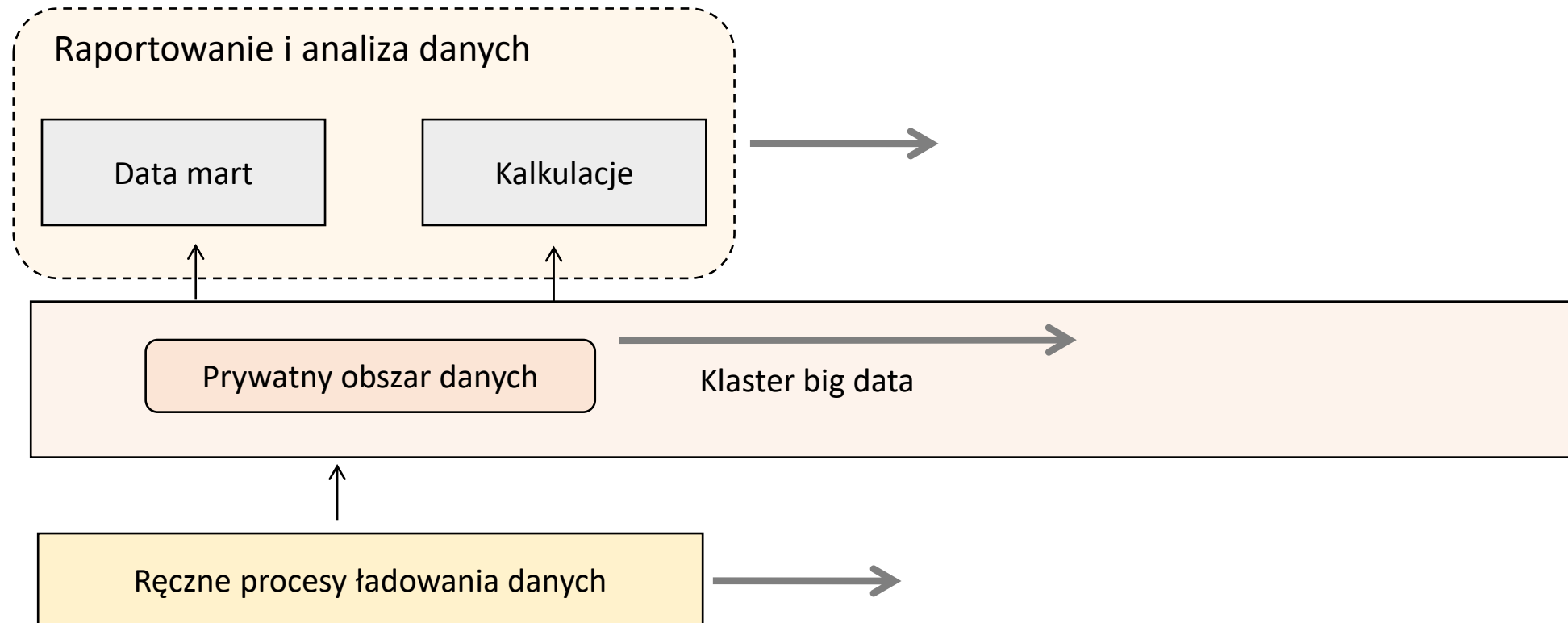
Przykład: big data jako repozytorium danych



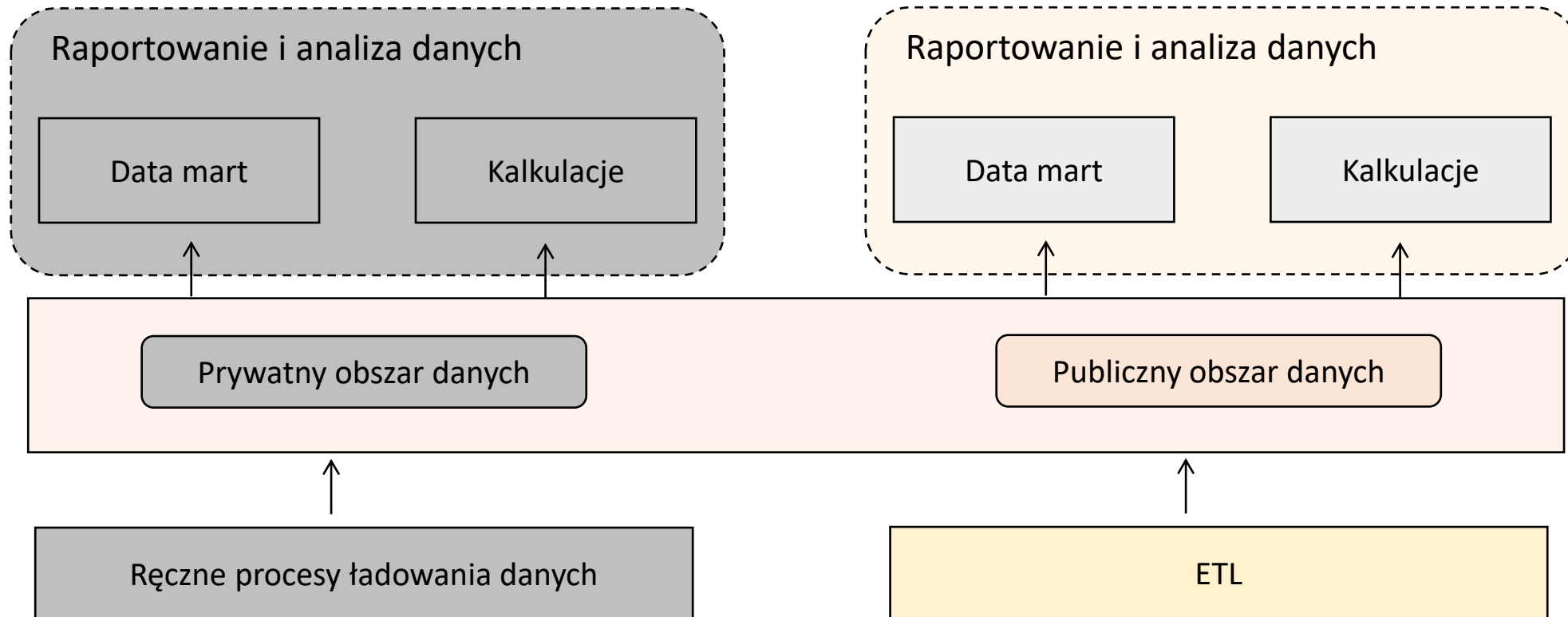
Przykład: środowisko self-service BI



Przykład: środowisko self-service BI



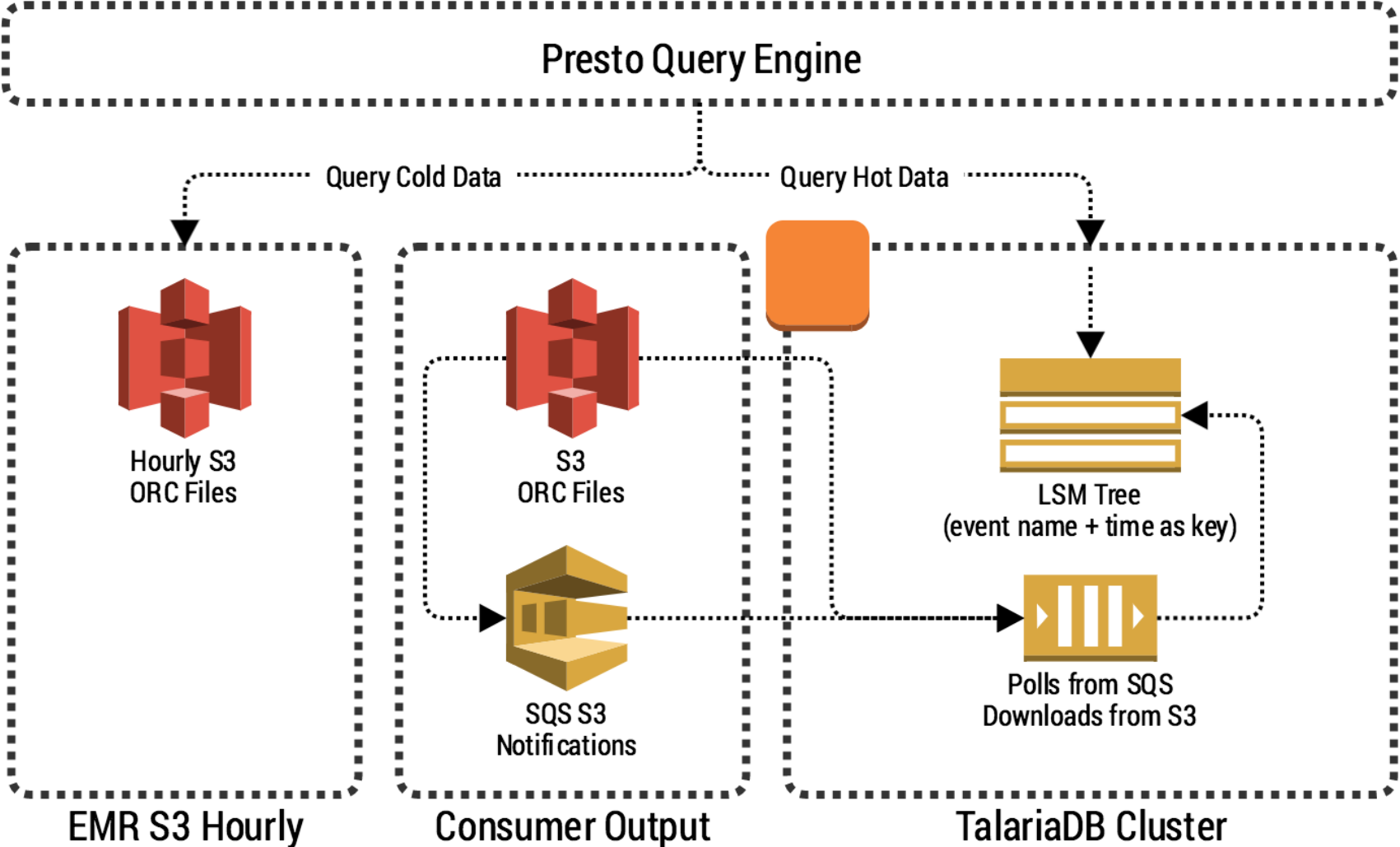
Przykład: środowisko self-service BI



Zagadnienia

- Uprawnienia
- Komunikacja HD – Big Data
- Metadane
- Redundancja

Presto



<https://engineering.grab.com/big-data-real-time-presto-talariadb>

Dziękuję za uwagę