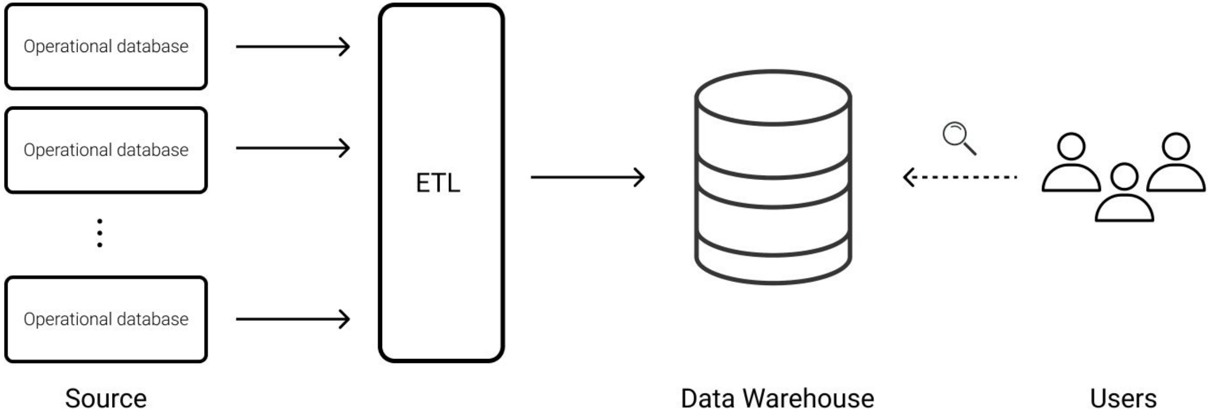
Consistency and Fault-Tolerance in Data Warehouses

Konzistentné dátové sklady a ich odolnosť voči chybám

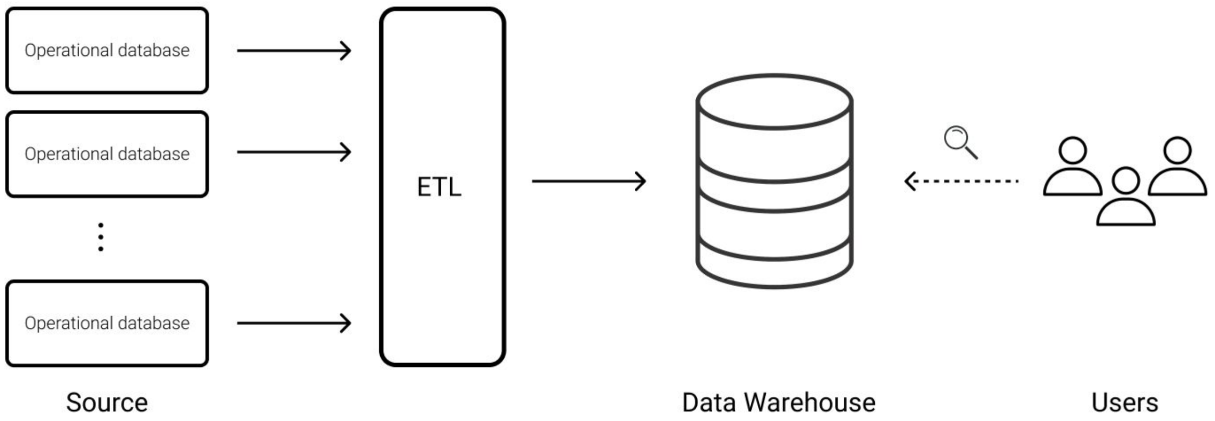
Radka Ďurčová
supervisor: Mgr. András Varga, PhD.



Big data
- computer clusters, distributed systems
- increased risk of failure

Objectives of a recovery strategy
- maintain data consistency
- performance

This thesis
- ETL process failures
- Dependency Analysis vs non-optimized approach



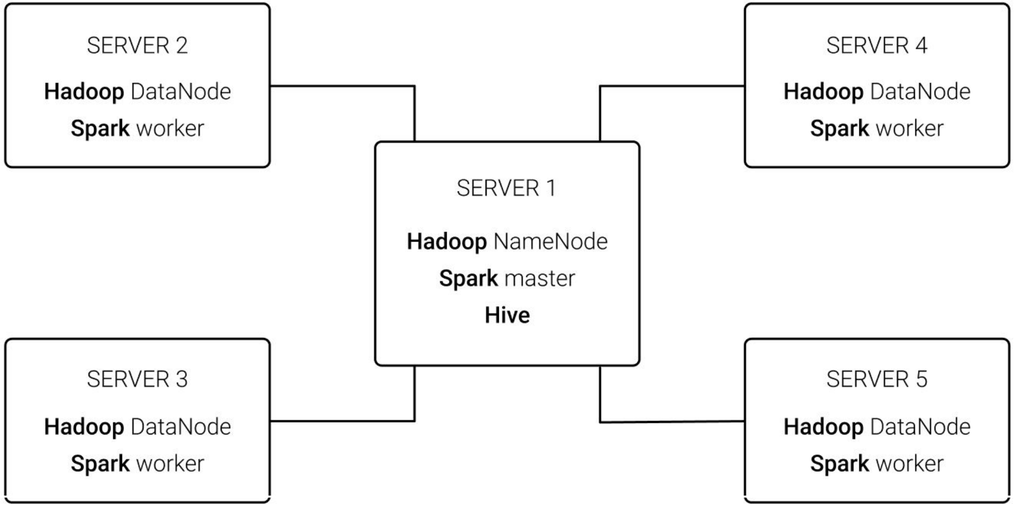
Data Warehouse

analytic database

ETL process

Extract, Transform, Load

Cluster



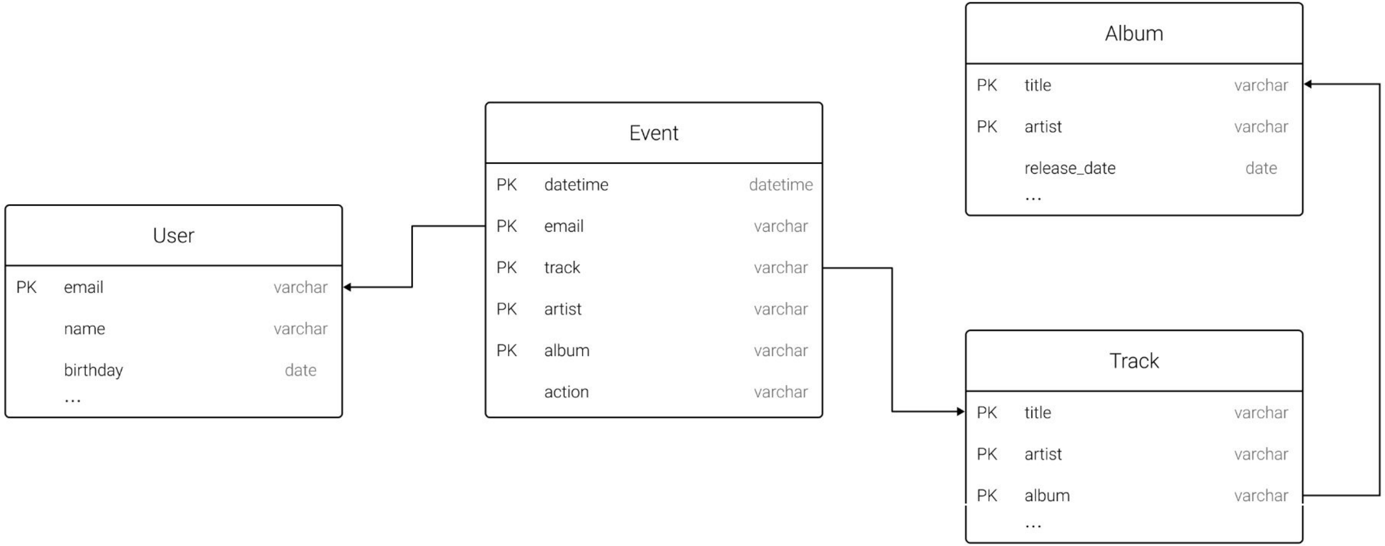
Data storage

- Hive
- Hadoop DFS

Data transformations

- Spark

Source system



Album

PK	title	varchar
PK	artist	varchar
	release_date	date
	...	

Event

PK	datetime	datetime
PK	email	varchar
PK	track	varchar
PK	artist	varchar
PK	album	varchar
	action	varchar

User

PK	email	varchar
	name	varchar
	birthday	date
	...	

Track

PK	title	varchar
PK	artist	varchar
PK	album	varchar
	...	

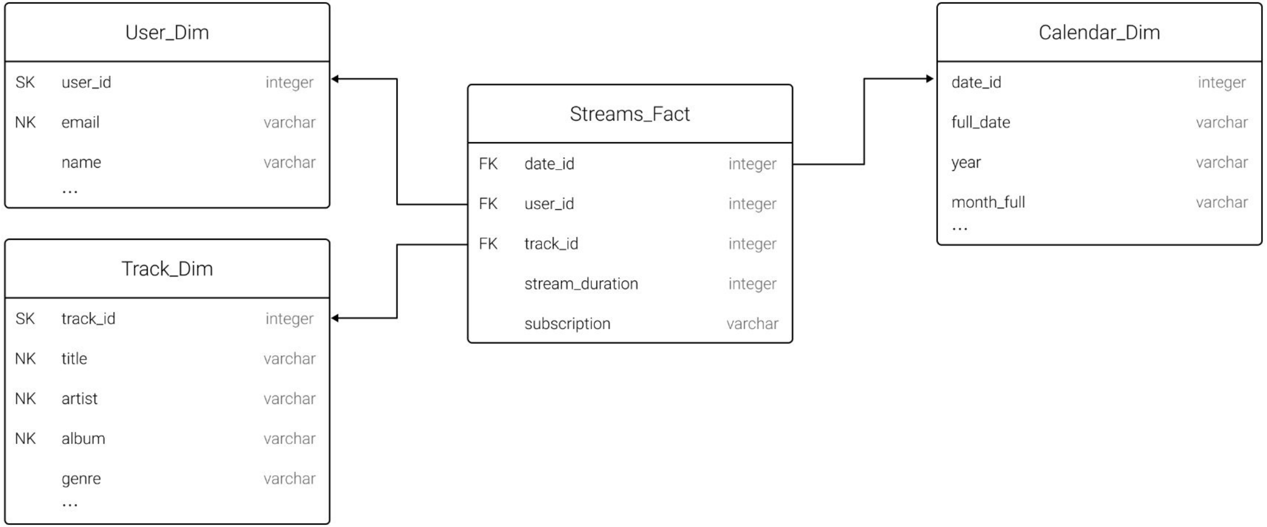
Fictional client

music streaming service

Source system

CSV files

Dimensional model



User_Dim		
SK	user_id	integer
NK	email	varchar
	name	varchar
	…	

Track_Dim		
SK	track_id	integer
NK	title	varchar
NK	artist	varchar
NK	album	varchar
	genre	varchar
	…	

Streams_Fact		
FK	date_id	integer
FK	user_id	integer
FK	track_id	integer
	stream_duration	integer
	subscription	varchar

Calendar_Dim	
date_id	integer
full_date	varchar
year	varchar
month_full	varchar
…	

Star schema

- fact tables - measurements
- dimensions - context

Methods

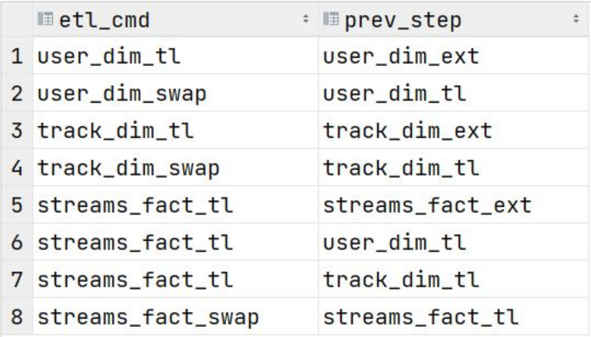
Naive approach

Dependency Analysis

3 stages: extract, transform-load, swap

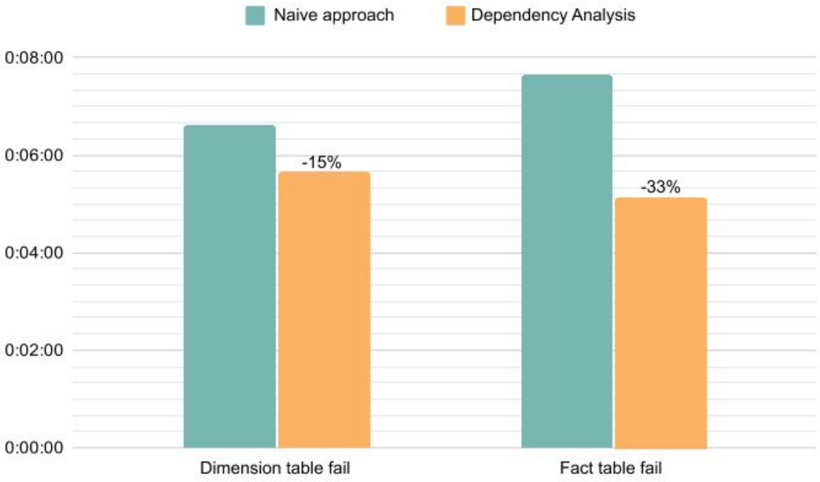
auxiliary tables

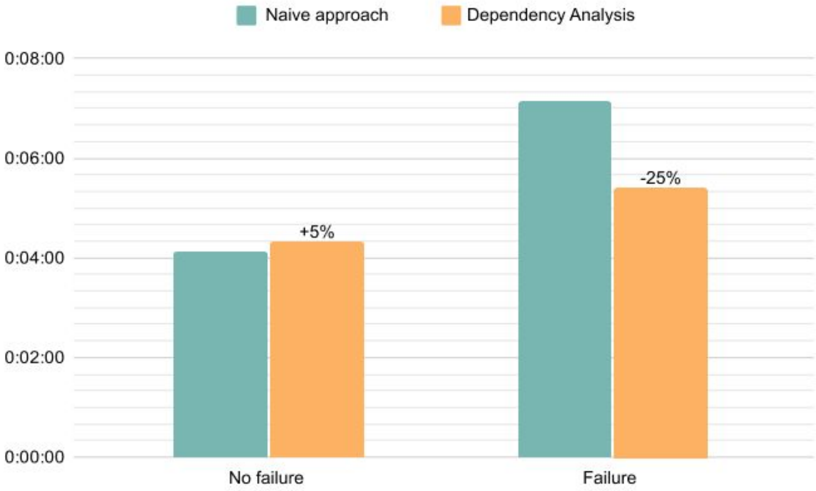
execution conditions, storing intermediate results



	etl_cmd	prev_step
1	user_dim_tl	user_dim_ext
2	user_dim_swap	user_dim_tl
3	track_dim_tl	track_dim_ext
4	track_dim_swap	track_dim_tl
5	streams_fact_tl	streams_fact_ext
6	streams_fact_tl	user_dim_tl
7	streams_fact_tl	track_dim_tl
8	streams_fact_swap	streams_fact_tl

Results





Dependency Analysis

- performance improvement
- minimal deceleration of a regular ETL run