



# The SQL Approach to Big Data

let them\* do the **hard** work:  
write **less** code, do **more** analysis  
have **more** fun!

Dr. Manolis M. Tsangaris,  
Head of Analytics, Performance Technologies

*manolis.tsangaris@performance.gr*

**Think Ahead.**

V1.4

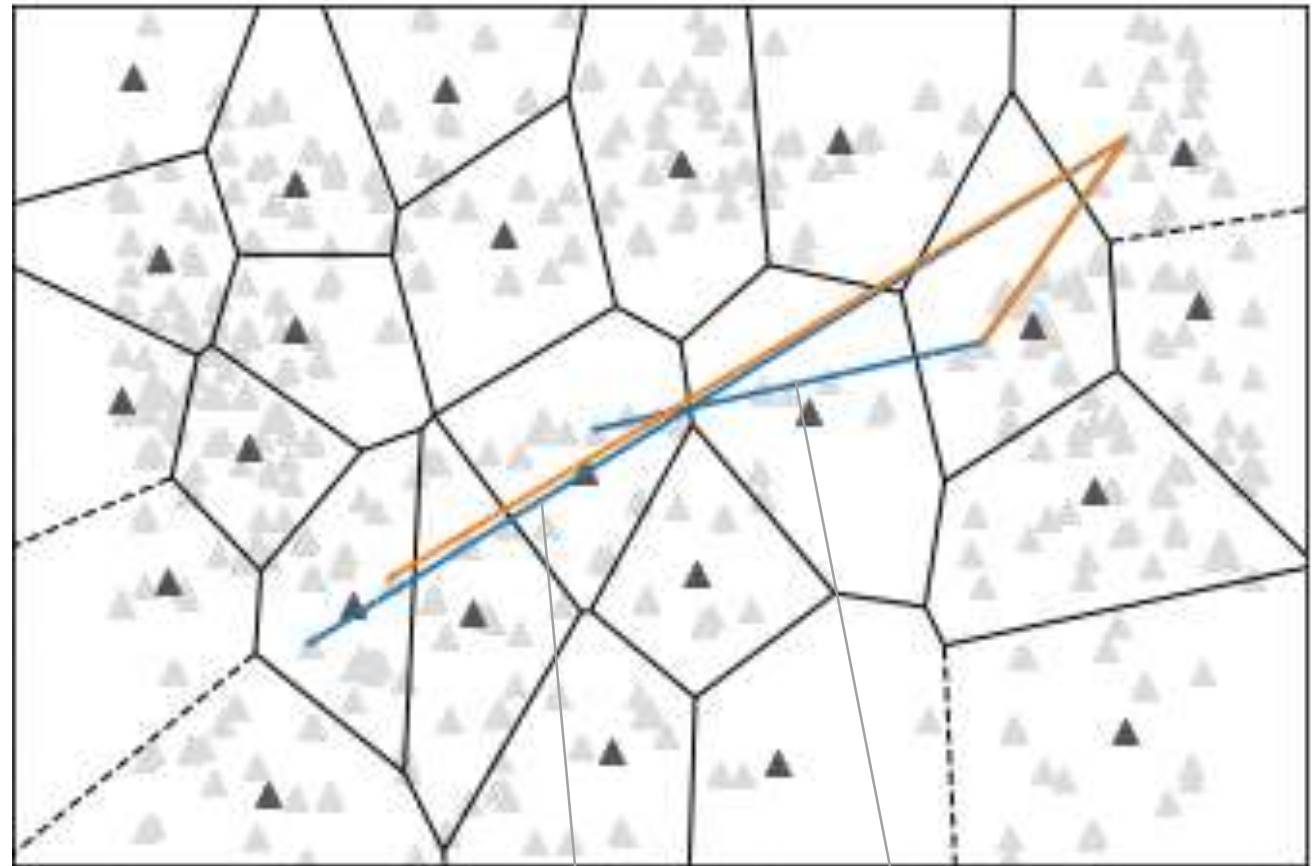
# A real world example ...

---

- Suppose you are a telecom company – you have 3M customers
- You want to help your customers find stolen phones
- You are sitting on tons of data:
  - Cell Tower Registrations:
    - **3 Million** active phones x 20 messages/hour = **1 Million** messages/second
    - **31 Trillion** messages per year
  - A thief steals your customer's phone
    - and keeps it for a while turned on along with their's
    - Thief's phone and stolen phone are traversing the network together
- Can I **find** thief's phone number?

# Let's map out the theft problem ...

- The mobile network consists of **cells**
- Cells cover **specific areas**
- The phones register
  - with the **best** cell tower every so often
- As the phone moves
  - **new cell** towers “register” the phone
- **Two** colocated phones may ...
  - Register to **different** cells



victim

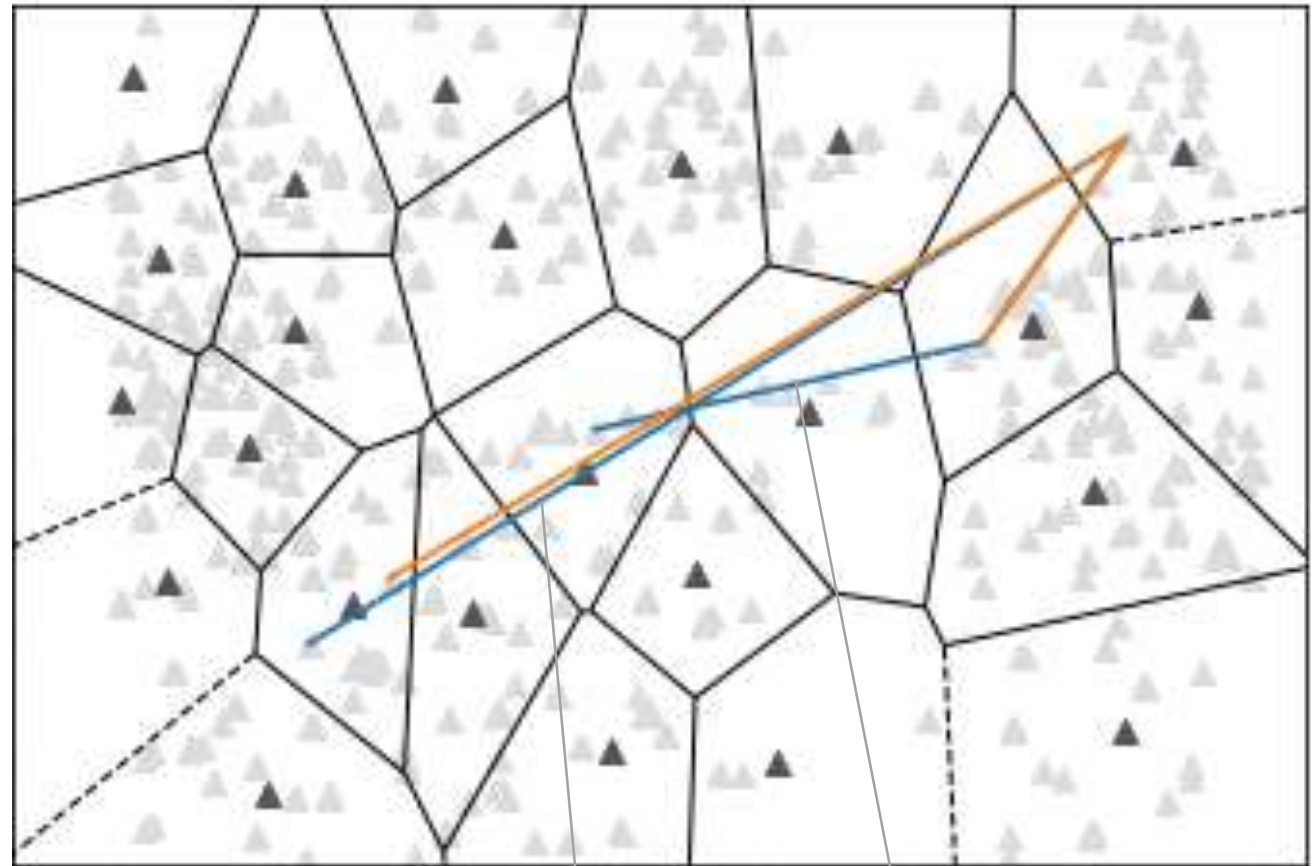
thief

**Think Ahead.**



# Let's map out the theft problem ...

- V (victim's phone) traverses the network (blue)
- T (thief's phone) traverses the network on a similar path (red)
- V path: 1,2,3,4,5,6,7,...
- T path: 9,11,3,4,5,6,7,...
- Thief has the same path as the victim after the point and time of theft



victim

thief



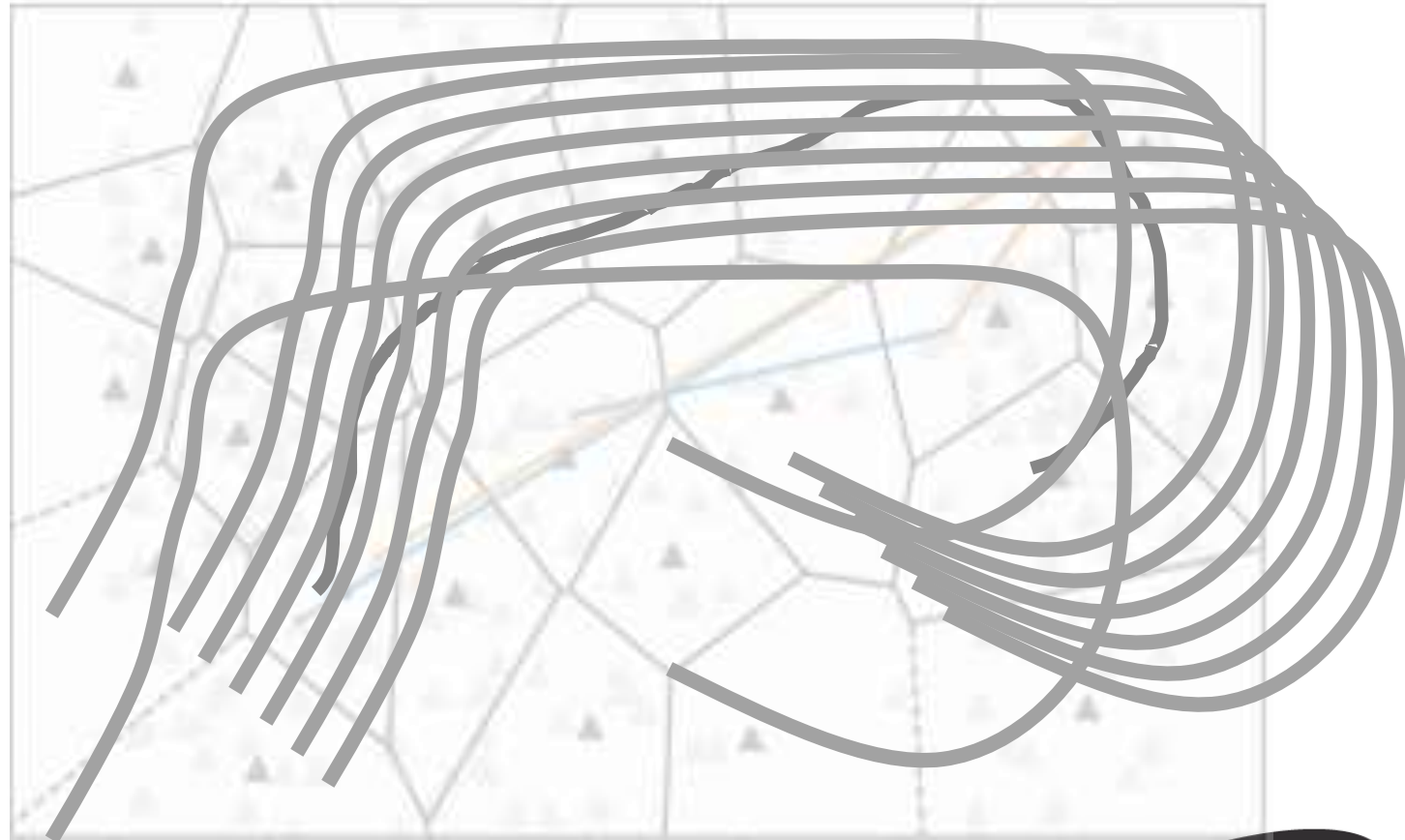
# The path is fuzzy .. the problem is hard

- The phones are heard from adjacent cells
- Multiple paths appear for each phone
  - not necessarily the same ones for each phone
- *While 3 million other phones*
  - *Keep moving!*



# The path is fuzzy .. the problem is hard

- The phones are heard from adjacent cells
- not necessarily the same ones for each phone
- other phones may follow similar paths
  - public transit or highway



# A (first) answer to this problem was found **quickly**

- It took an **SQL Analytical Platform** and an experienced **data scientist**
- Implementation:
  - **5 hours** to implement the first running version
  - **1 more day** to validate it
  - **1 more day** to go to production!
  - **Less than 200 lines of SQL code**
- Correctness:
  - 95% (in validation - confirmed with known cases)
- Performance:
  - Finds the potential thief in about **5 minutes**
- Velocity:
  - Could still deal with live data coming in @ 1M messages/second
- Bonus:
  - It could even track many thief cases at the same time



# The (large telecom) customer

---

- ... made up his mind
- And ordered the SQL based analytics platform
  - and made even more money from network data monetization
- The Happy End!
  - and this is not a children story but a real customer case
- *I am sure you find similarity of this problem to many other business problems*
  - *Correlating customer behaviors, locating customer journeys, detecting fraud, ...*



**Think Ahead.**



# The Anatomy of the solution

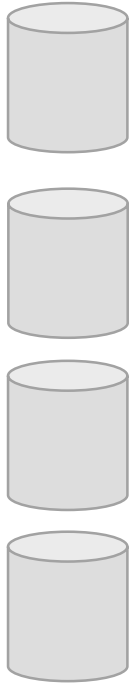
---

**Think Ahead.**



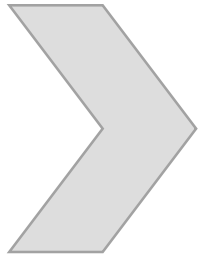
# The platform

---



## Velocity:

Ingestion @ 1M records/second



SQL based Analytics Platform

**Volume:** about **40 Terabytes/year** (10 TB compressed)

## Variety:

- order of magnitude **different** behavior of mobile customers
- **Evolving** network and mobile phone base

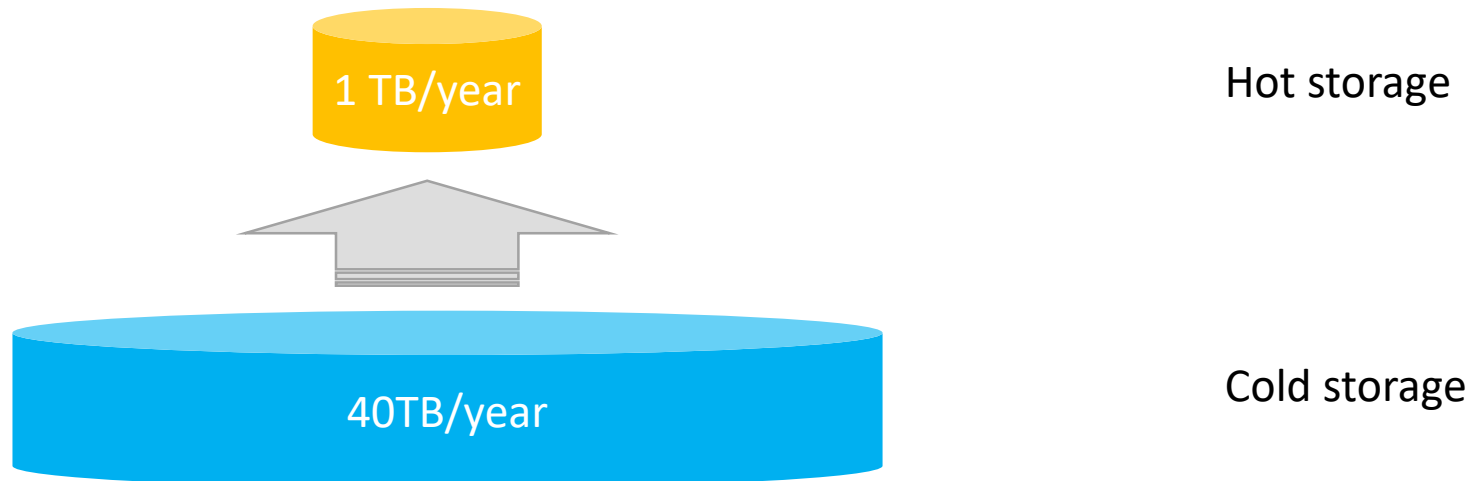
**Think Ahead.**



# The modeling

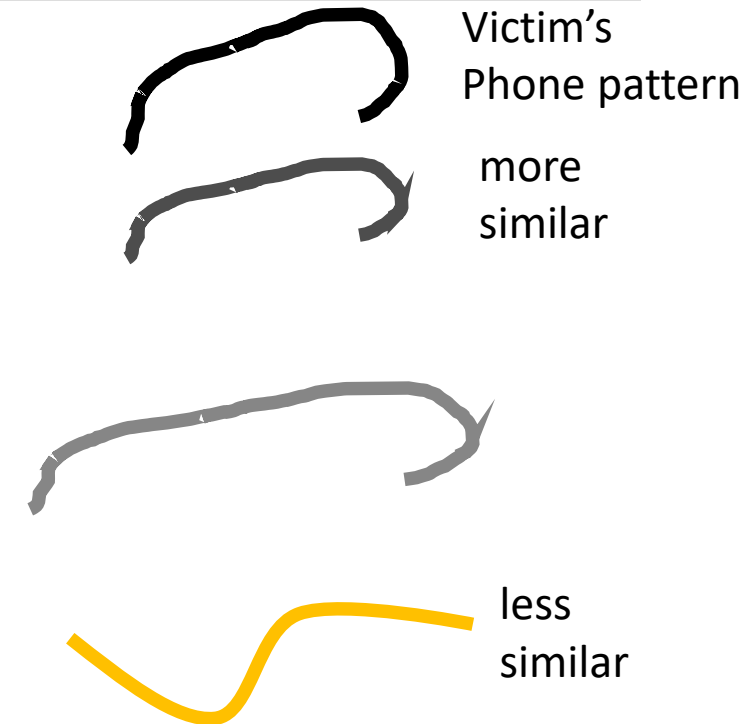
---

- Compress the events of stationary customers
  - A as much as 100 lines of SQL code with timeseries processing
  - Achieved 4:1 reduction of data



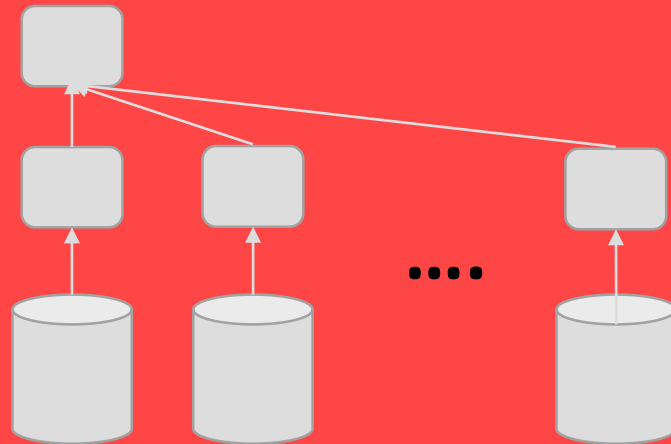
# The builtin analytics

- **Phone trajectory:** the path of a phone across the network
  - Space (location of cell towers)
  - Time (time of the registration)
- **Similarity** of phone trajectories based on their “paths”
  - Two dimensional similarity function
- Find all phones that are “**similar**” to the victim phone
  - Starting at the time and location of the theft (reported by the victim)
- “**calibrate**” and **evaluate** the similarity function
  - based on known trajectories of theft
- All of the above are expressed using **Analytical SQL functions**
  - **Run distributed over the Analytical SQL Engine**



# The efficient execution using parallelism

## Parallel Distributed Execution



SQL based Analytics Platform

- **Scaleup:**
  - The platform can scale horizontally
- **Elasticity:**
  - The workload can enjoy limited or full parallelism
- **Devops:**
  - The platform can grow with no downtime

Think Ahead.



So what was this SQL Analytics Platform?

VERTICA

# Vertica SQL Analytical Platform is:

---

- The leading on Premises AND on Cloud analytical platform
- Parallel & Distributed - shared nothing or common storage
- N+1 scalable
- SQL based with extensions
- Can remap and run Python based ML code in core



# How about us (Performance Technologies)?

---

- We have designed, deployed and delivered
  - **One of the Largest Based Big Data Platforms in Greece**
  - In a major telecom provider
- Facts:
  - **Ingests** more than 1TB/day data from more than 450 sources
  - **Grows** at a rate of 5TB/month
  - Keeps in (Hot and Cold) storage more than 30TB of **historical data**
  - **Executes** around 5 million analytical SQL queries per day
  - Has undergone **scale out** expansion three times with no downtime
  - Supports **mixed workloads** of ML, Exploration, and Production workloads
  - And does this at about 1/3 to 1/5 of the **total cost** of ownership\*

# Plus some more benefits

---

- **Speed of delivery:**
  - Like the 'thief case', new applications were:
    - prototyped in hours\*
    - Fully developed in days
    - Go to production in a few weeks
- **Ease of use:**
  - Complex data relationships
    - Were expressed, evaluated, optimized
    - By a wide audience – Data Democratization
  - SQL Table access
    - made it very convenient to serve analysts and developers\*
- **Stability:**
  - Application performs just scales up/down
    - With increasing workload
    - With increasing resources

*\* It also kept DSs and dbadmins happy!*



# Next Steps

---

- We love to talk with you on how to:
  - .. let **VERTICA** do the hard work for you
    - While you write less code
    - Do more analysis
    - And have more fun doing it!
- We are Performance Technologies ([www.performance.gr](http://www.performance.gr))
  - We are a **trusted** partner for organizations that seek to redefine and reinvent themselves through **digital technologies**.
  - We provide **products, services and solutions** that transform traditional businesses into digital leaders.
  - We help businesses **grow** through the most effective use of technology

**Think Ahead.**



# PERFORMANCE AT A GLANCE

**200+**  
**TRANSFORMATION**  
WORKSHOPS **DELIVERED**

## TECHNOLOGY

- WORLD-CLASS SKILLS
- KEY VENDOR ALLIANCES
- FOCUS ON EXECUTION
- CUSTOMER CENTRIC

## SOLUTION AREAS

- CLOUD STRATEGY
- BUSINESS WORKFLOWS
- ANALYTICS & BIG DATA
- CYBERSECURITY & RISK

**TRUSTED**  
BY **90%** OF  
ENTERPRISES  
IN GREECE

**80+ CERTIFIED**  
**PROFESSIONALS**

**450+ CERTIFICATIONS**

**AUTHORIZED AWS, AZURE**  
& **GOOGLE CLOUD PARTNERS**

We are first and foremost driven by a deep-rooted engineering ethos that drives us to deliver the best possible solutions at any given time!