Analytics in 2024 can hurt your database

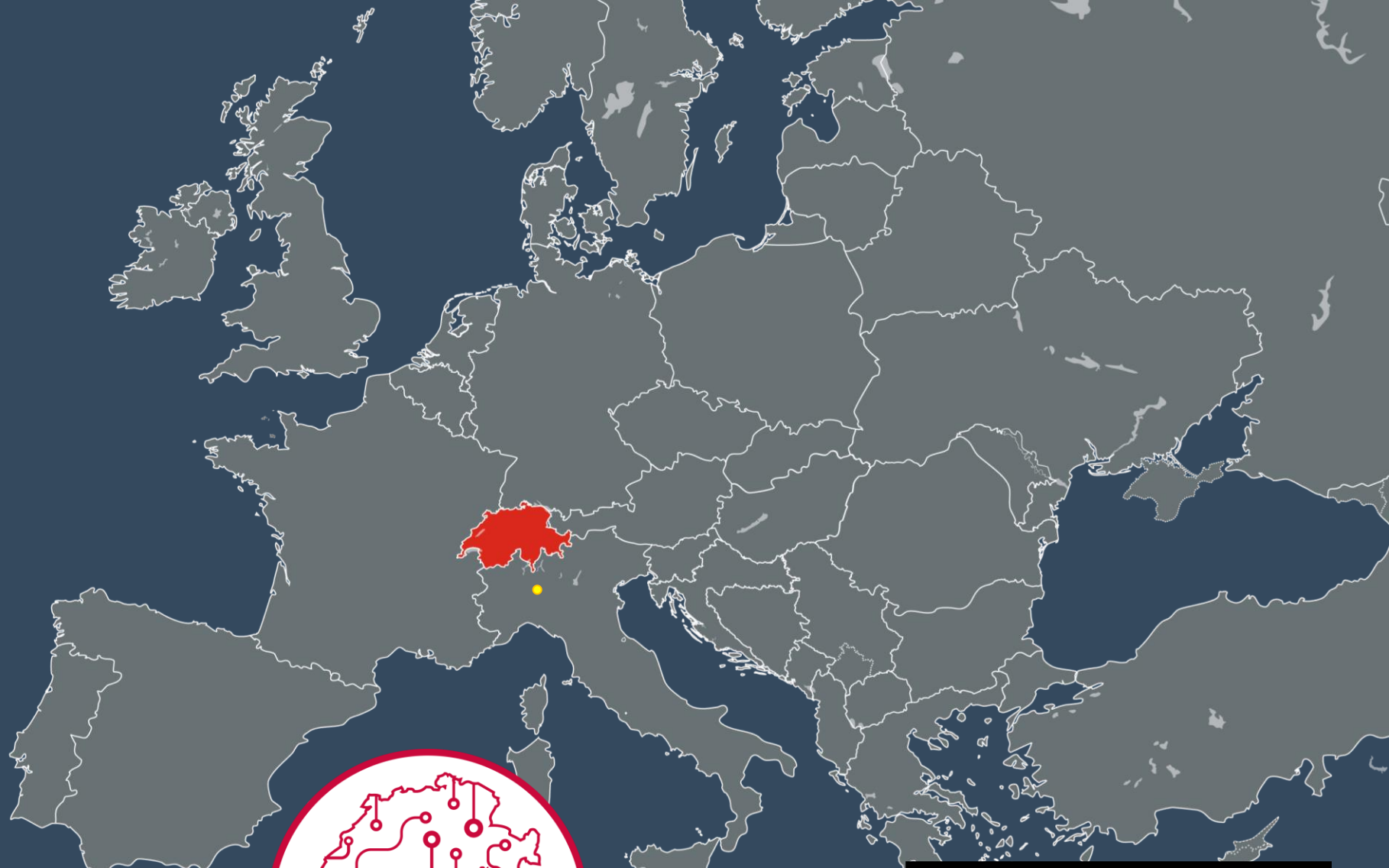
In new, unexpected, ways

October 24, 2024

@G_Ceresa
W: www.datalysis.ch
E: info@datalysis.ch

Gianni Ceresa

Working with ***data***,
Business Analytics
and EPM tools
for more than
15 years



Oracle ACE
Director

SWISS OUC

SYM42

Oracle ACE

430+ technical experts helping peers globally

The **Oracle ACE Program** recognizes and rewards community members for their technical and community contributions to the Oracle community

3 membership tiers

Oracle ACE Director

Oracle ACE Pro

Oracle ACE Associate

For more details on Oracle ACE Program:
ace.oracle.com

Oracle ACE

Nominate

yourself or someone you know:

ace.oracle.com/nominate

Connect: **aceprogram_ww@oracle.com** Facebook.com/OracleACEs @oracleace Oracle ACE Program Group

DISCLAIMER

I'm an impostor!

DISCLAIMER

I'm an impostor!

I'm not a DBA, I never claimed to be one, I don't aspire at being one. No, it isn't you, it's me...

But I happen to be one of those hammering your database with new weird workloads that you didn't expect, nor imagine.

Mostly because I can, sometime because I have to...

It is based on all the things I saw (and did) in the past 15 years... And this makes me feel old...

Analytics in 2024 can hurt your database

In new, unexpected, ways

A step back before to jump forward

The current situation, the new unexpected ways your database is being used and abused, comes in opposition to the past.

A relational database is still mostly the same thing as it was many years ago: more optimizations, but still the same relational model of data.

The DBA role has been defined by what the job was when it started being a thing years ago.

Like (almost) everything, they both have to evolve with time, they must keep up with the changes in the field:

- Natural evolution, improvements
- Buzzword evolution, temporary trends

Reporting:
Known, simple, “static”

Once upon a time there was reporting...

Reporting is very old.

The idea is some predefined queries, executed based on a schedule, producing an output (Excel file?) the users could further analyse locally on their own.

Mostly done in the database directly.

Or via some tools rendering reports with the retrieved data.

Once upon a time there was reporting...

Because the queries are known and their execution schedule as well, it's easy to tune the query, to make sure the database has everything to execute the query in the best conditions (indexes, partitions etc.), to schedule other tasks to not overlap too much.

DBAs easily can have a pro-active role in this, because the context and queries are known, therefore they can prevent issues (with increasing volume of data etc.).

Old style reporting can still be a thing

Order Entry Data Model

Home | Catalog | Favorites | Dashboards | Create | Open

Validate | Manage Private Data Sources | View Data | Create Report

Data Model

- Properties
  - Data Sets
    - QUERY
  - Event Triggers
  - Flexfields
  - List of Values
    - order-list
    - customer-list
  - Parameters
    - cstid
    - ordid
  - Bursting

Diagram | Structure | Data | Code

Global Level Functions

Drop here for aggregate function

ROW

QUANTITY_ON_HAND	#
QUANTITY	#
STREET_ADDRESS	A
POSTAL_CODE	A
CITY	A
STATE_PROVINCE	A
COUNTRY_NAME	A
CREDIT_LIMIT	#E
CUST_EMAIL	A
PRIMARY_PHONE_NUMBER	A
CUSTOMER_ID	#
ORDER_TOTAL	#E
ORDER_STATUS	#
ORDER_DATE	A

Edit Data Set - QUERY

* Name: QUERY

* Data Source: Oracle BI EE (Default)

* Type of SQL: Standard SQL

* SQL Query — Query Builder

```
SELECT nvl(wh.quantity_on_hand,0) quantity_on_hand,
       order_items.quantity,
       cv.street_address, cv.postal_code,
       cv.city, cv.state_province,
       cv.country_name, cv.credit_limit,
       cv.cust_email, cv.primary_phone_number,
       cv.customer_id,
       ov.order_total, ov.order_status,
       to_char(ov.order_date,'YYYY-MM-DD') ORDER_DATE, order_items.unit_price,
       product_information.product_name,
       product_information.product_description,
       (cv.cust_first_name || ' ' || cv.cust_last_name
       ) customer_name,
       (employees.first_name || ' ' || employees.last_name) sales_mgr,
       order_items.order_id, order_items.line_item_id
```

Generate Explain Plan | OK | Canc

Old style reporting can still be a thing

Order Entry Data Model

Home | Catalog | Favorites | Dashboards | Create | Open

Validate | Manage Private Data Sources | View Data | Create Report

Data Model

- Properties
  - Data Sets
    - QUERY
  - Event Triggers
  - Flexfields
  - List of Values
    - order-list
    - customer-list
  - Parameters
    - cstid
    - ordid
  - Bursting

Diagram | Structure | Data | Code

Table View | Output

	XML View			Business View	
Data Source	XML Tag Name	Sorting	Value If Null	Display Name	Data Type
Report Data					
Data Structure	ROWSET				
QUERY	ROW				
QUANTITY_ON_HAND	QUANTITY_ON_HAND			QUANTITY_ON_HAND	#
QUANTITY	QUANTITY			QUANTITY	#
STREET_ADDRESS	STREET_ADDRESS			STREET_ADDRESS	A
POSTAL_CODE	POSTAL_CODE			POSTAL_CODE	A
CITY	CITY			CITY	A
STATE_PROVINCE	STATE_PROVINCE			STATE_PROVINCE	A
COUNTRY_NAME	COUNTRY_NAME			COUNTRY_NAME	A
CREDIT_LIMIT	CREDIT_LIMIT			CREDIT_LIMIT	#E
CUST_EMAIL	CUST_EMAIL			CUST_EMAIL	A

Old style reporting can still be a thing

Customer Orders

Home Catalog Favorites Dashboards Create Open

Data Model Order Entry Data Model

Parameters Properties View Report

View Thumbnails | View a list

Add New Layout

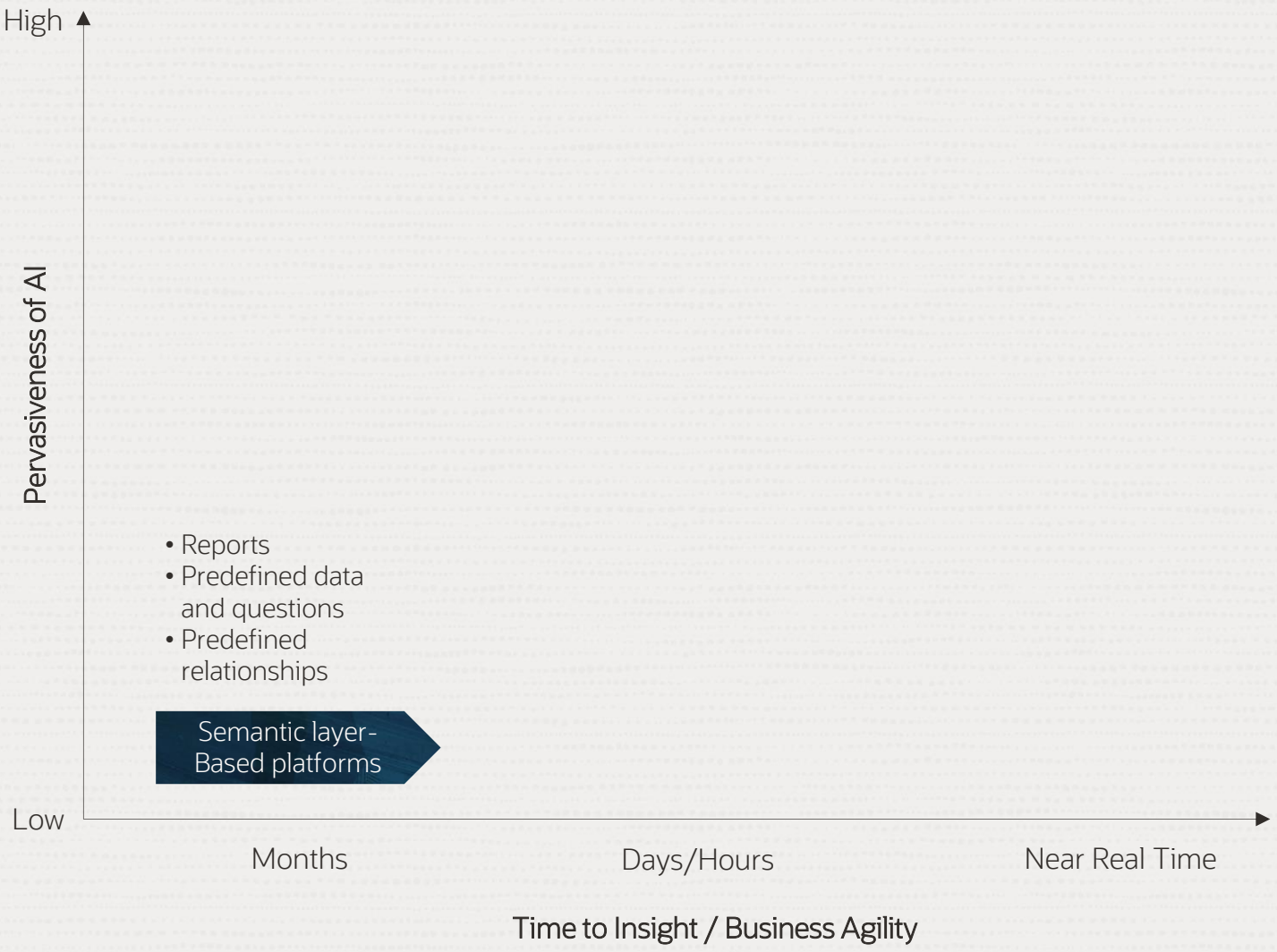
Customer Invoice
Edit | Properties | Delete

Customer Dunning
Edit | Properties | Delete

Customer Orders
Edit | Properties | Delete

Business Intelligence:
Centrally owned, under control

Timeline of Innovation Points in the Analytics Market



Source: Gartner 2021

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

Where are the skills ?

Only few resources hold technical skills to execute Analytics…

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

Oracle Analytics

Converged Analytics Platform for all Personas, Workloads and Data

Data Engineers

Governed Analytics	
Dashboards	Distributed Pixel-Perfect Reports
Semantic Models	Query Federation
Briefing Books	Data Export

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

The good old days of Business Intelligence

Business Intelligence is more “advanced” than reporting. It tried to tell a story, to focus on specific topics and display a complete overview of the various metrics, KPIs and data, using charts more than simple tables.

The BICC (Business Intelligence Competency Center) was in charge to collect the business needs, build the required dashboards and analyses, and deliver them to users for consumption (read-only).

Data is modelled, queries are written, by a small number of skilled people.

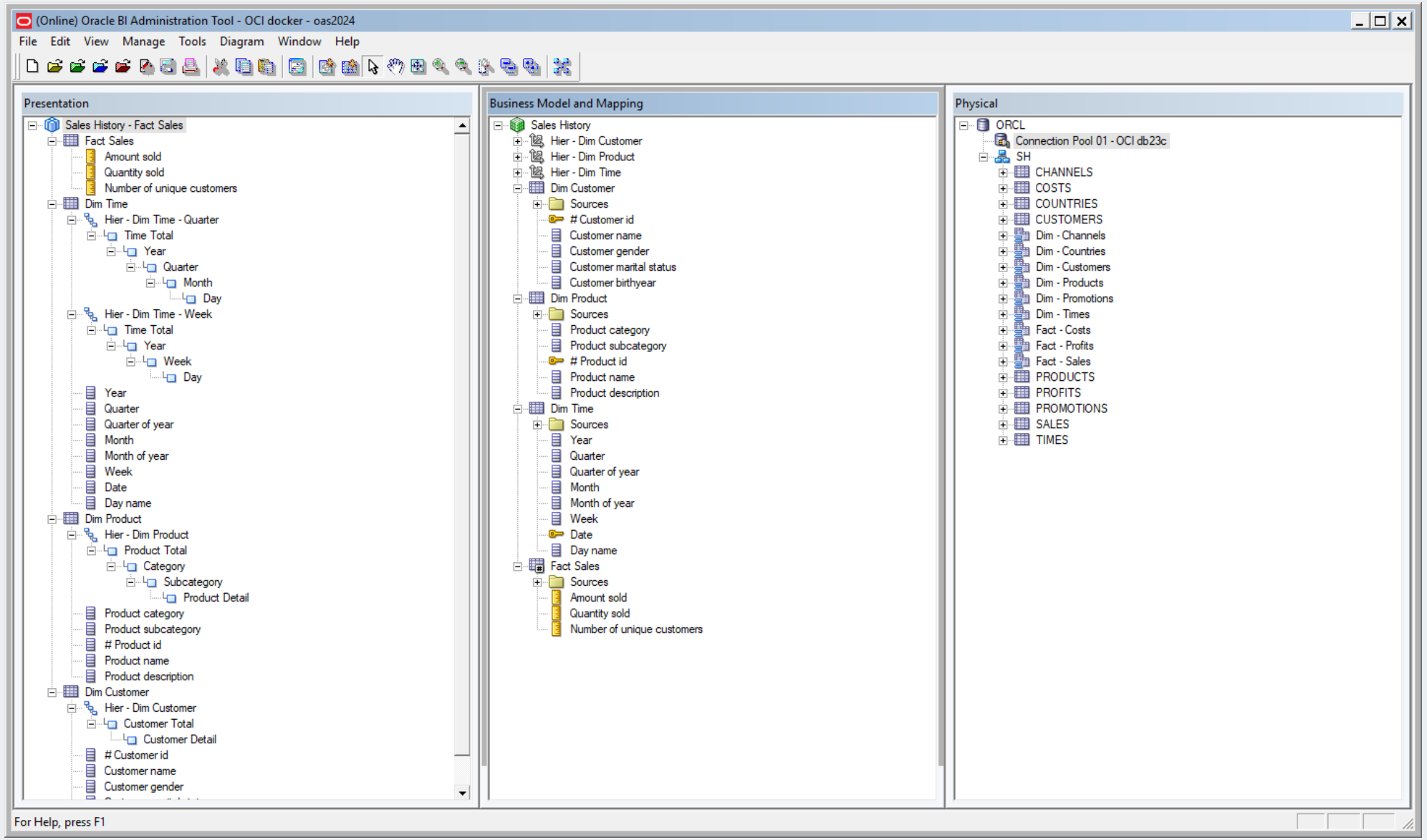
Users have a limited impact on the generated queries, at best setting filters to only see a subset of the whole data.

The good old days of Business Intelligence

In Business Intelligence, it is still possible to have a pro-active interaction between the BICC people and the database people.

Not just a fixed list of queries, but all the queries generated by dashboards are predictable and can be tuned. Data models can be adapted to support the BI needs instead of requiring complex, suboptimal, modelling.

Oracle Business Intelligence (Siebel and nQuire before, Oracle Analytics after)

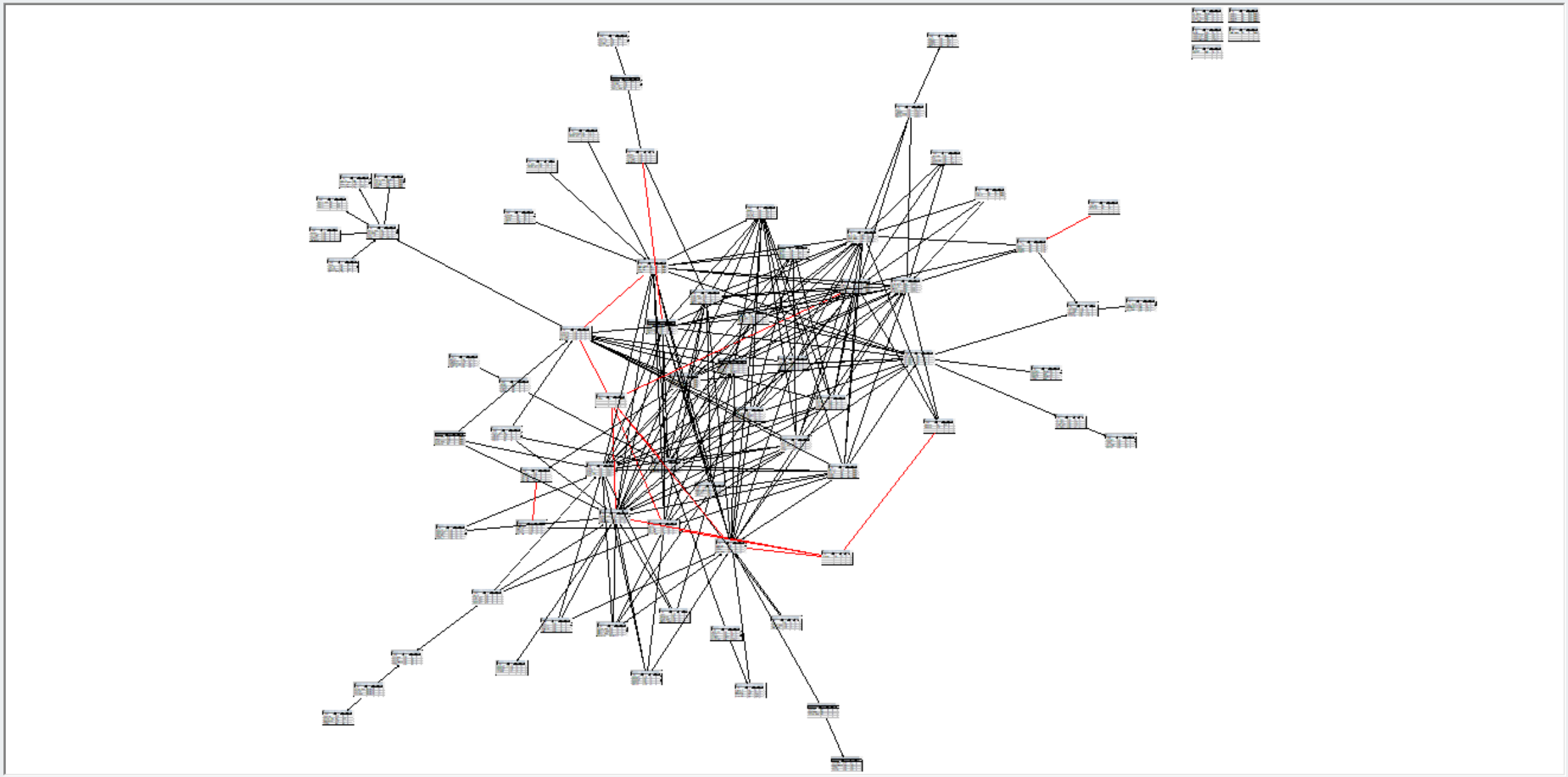


Oracle Business Intelligence (Siebel and nQuire before, Oracle Analytics after)

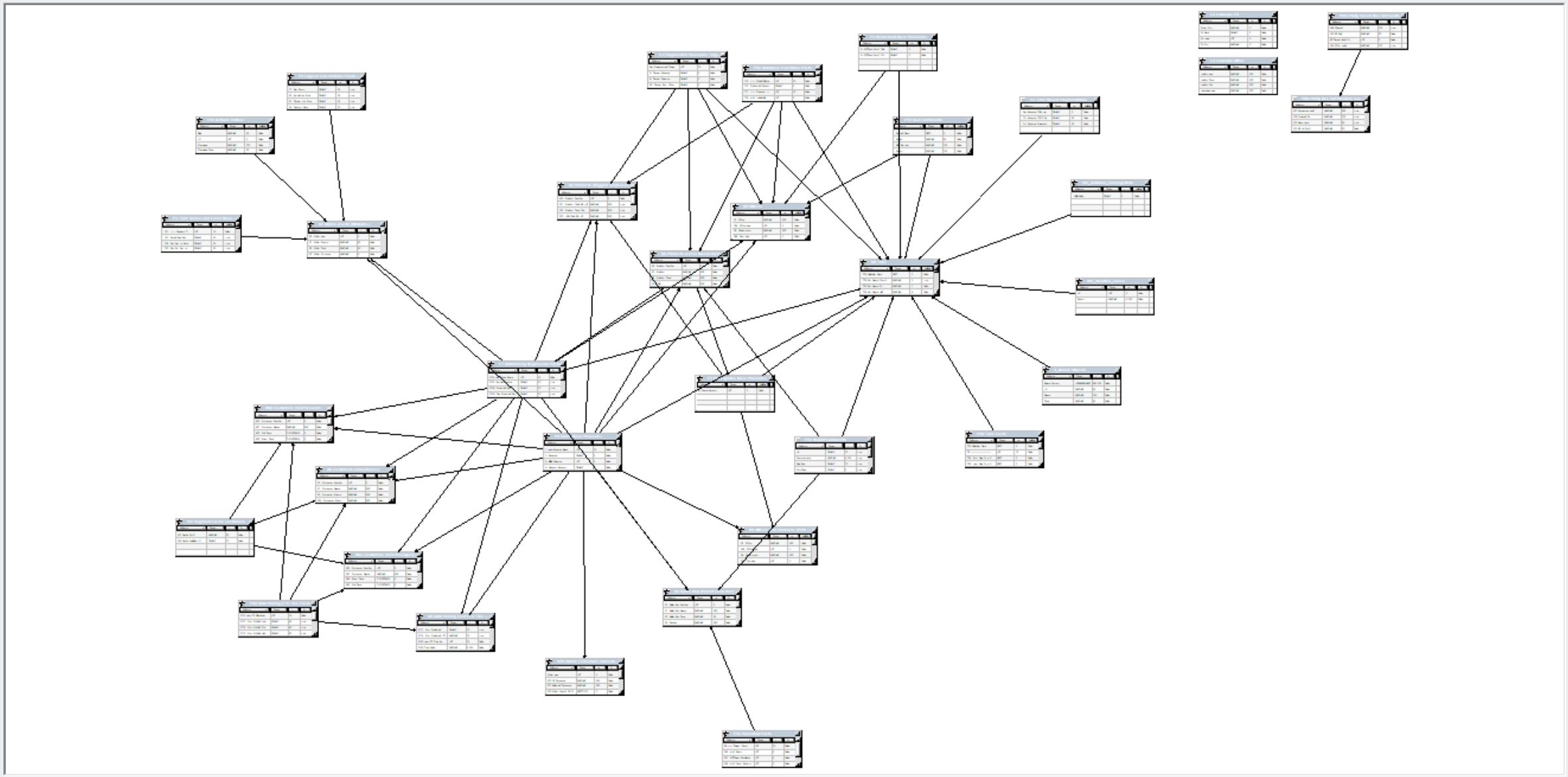
Queries generation can be optimized by using all the features available in the products:

- Pre-aggregated tables at higher levels
- Using multiple physical objects as one (for example for sales, if current year sales and historical sales are in different tables)
- Fixing granularity at which data is available
- Pre-filtering data as required avoiding queries on millions and millions of rows
    - For example: when opening a dashboard, request filters before to execute a query

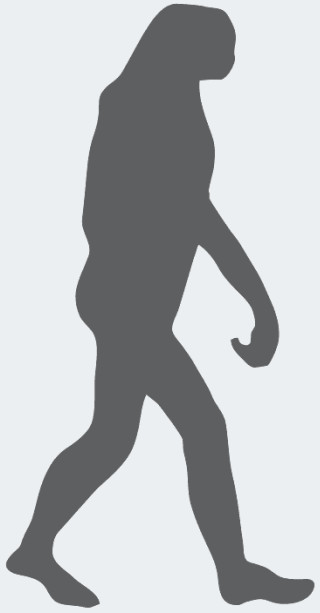
Oracle Business Intelligence (Siebel and nQuire before, Oracle Analytics after)



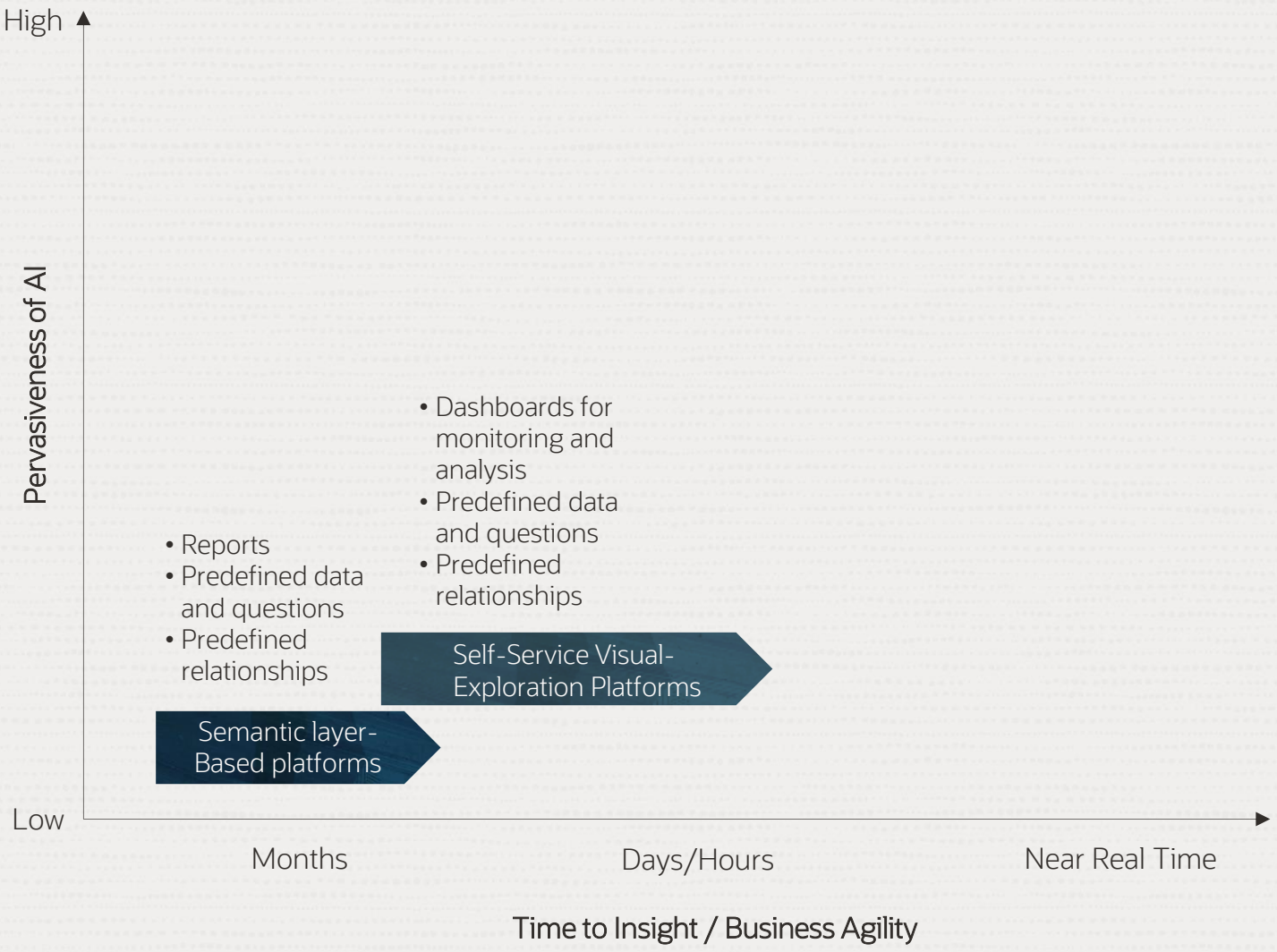
Oracle Business Intelligence (Siebel and nQuire before, Oracle Analytics after)



Self-Service Analytics:
Some form of control, at the beginning…



Timeline of Innovation Points in the Analytics Market



Source: Gartner 2021

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

Self-Service Analytics : a first step of freedom for the users

Business Intelligence made by the BICC team has a major issue: time

- It does take time to the small BICC team to deliver all the requirements of all the business units in a company.

Users want more freedom, they want to be able to make their own analyses.

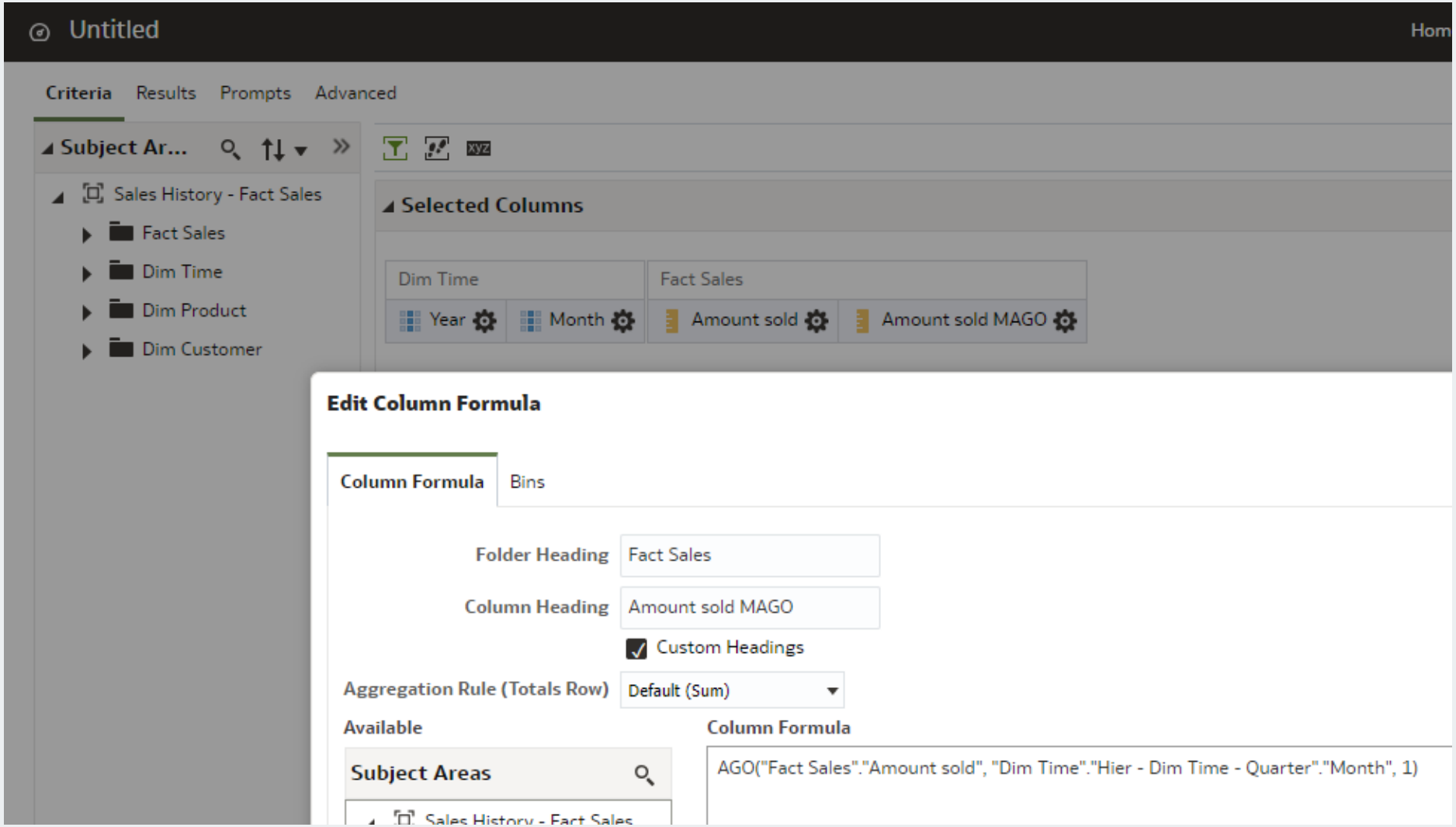
A role of "author" is introduced to allow a subset of users to not be simple consumers, but to build their own analyses and dashboards on top of the prebuilt model provided by BICC.

There is still some control because the metadata model is under control, but users are free to enter their own expressions, their own formulas in columns. This could lead to very weird queries being generated...

Self-Service downsides: users write their own analyses

A simple example:

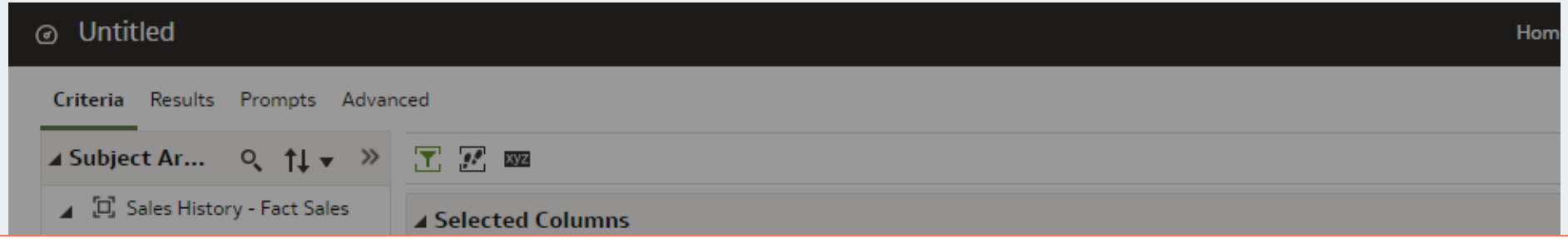
- Amount sold, Amount sold 1 month ago by year and month



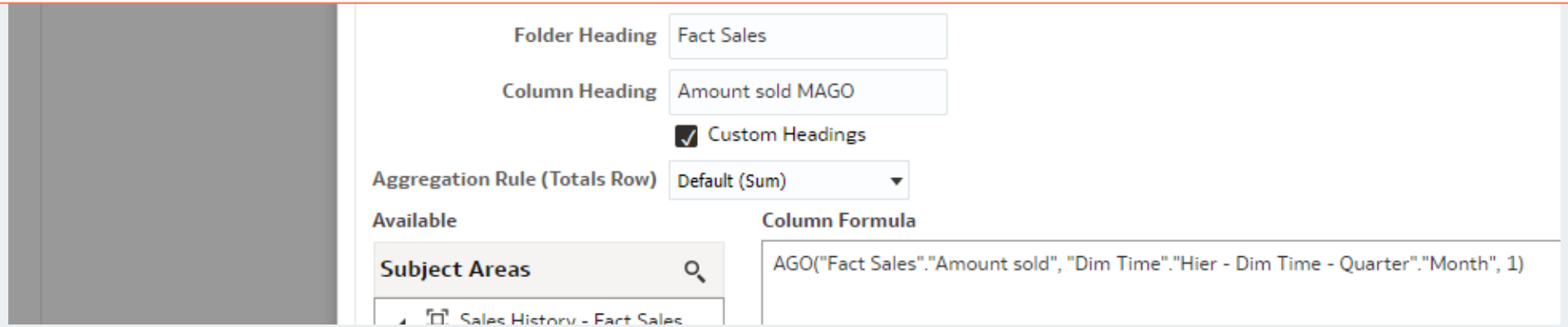
Self-Service downsides: users write their own analyses

A simple example:

- Amount sold, Amount sold 1 month ago by year and month



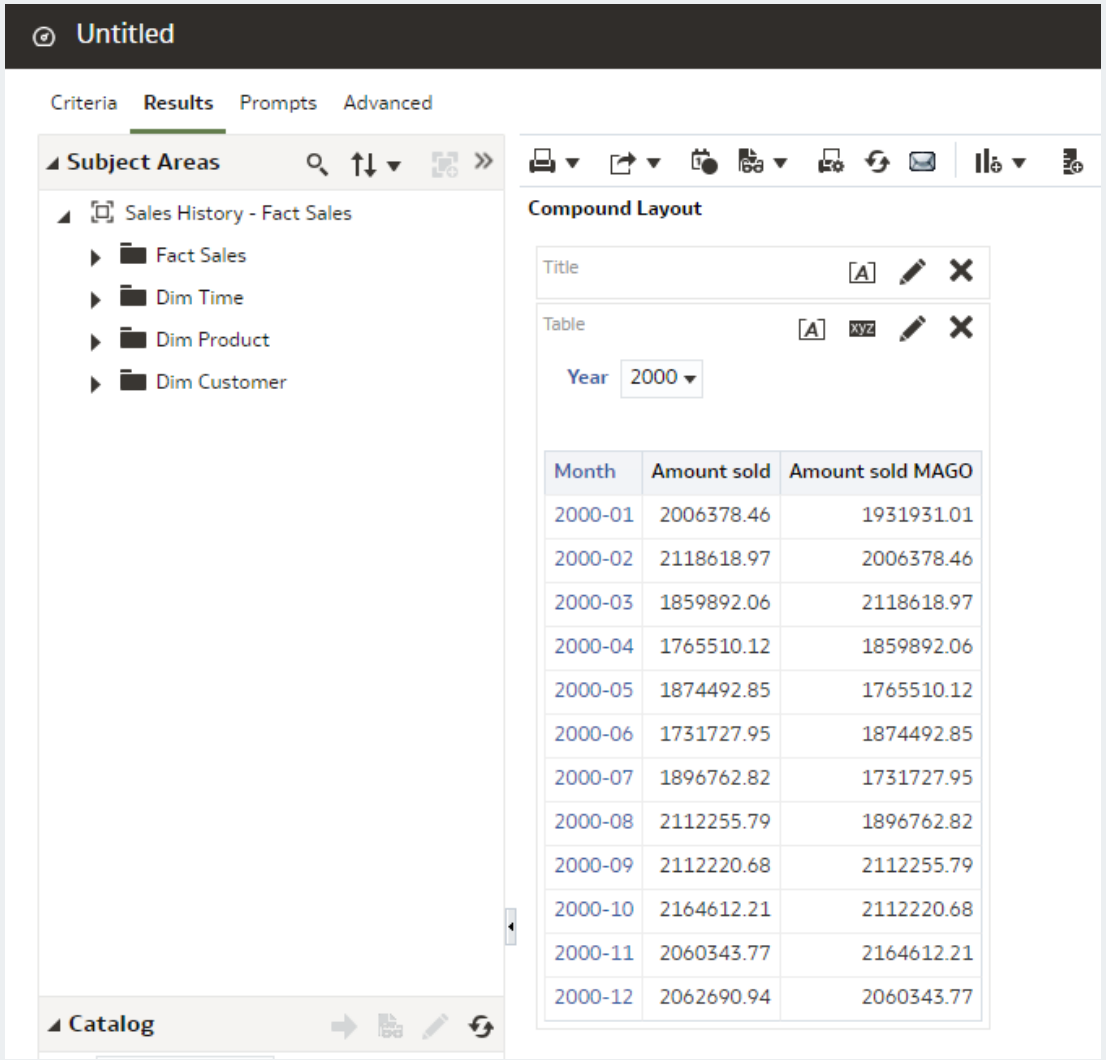

```
Logical SQL:

SELECT
   0 s_0,
   "Sales History - Fact Sales"."Dim Time"."Month" s_1,
   "Sales History - Fact Sales"."Dim Time"."Year" s_2,
   "Sales History - Fact Sales"."Fact Sales"."Amount sold" s_3,
   AGO("Sales History - Fact Sales"."Fact Sales"."Amount sold","Sales History - Fact Sales"."Dim Time"."Hier - Dim Time - Quarter"."Month",1) s_4
FROM "Sales History - Fact Sales"
ORDER BY 3 ASC NULLS LAST, 2 ASC NULLS LAST
FETCH FIRST 65001 ROWS ONLY
```



Self-Service downsides: users write their own analyses

A simple example:

- Amount sold, Amount sold 1 month ago by year and month



Month	Amount sold	Amount sold MAGO
2000-01	2006378.46	1931931.01
2000-02	2118618.97	2006378.46
2000-03	1859892.06	2118618.97
2000-04	1765510.12	1859892.06
2000-05	1874492.85	1765510.12
2000-06	1731727.95	1874492.85
2000-07	1896762.82	1731727.95
2000-08	2112255.79	1896762.82
2000-09	2112220.68	2112255.79
2000-10	2164612.21	2112220.68
2000-11	2060343.77	2164612.21
2000-12	2062690.94	2060343.77

Self-Service downsides: users write their own analyses

Physical SQL:

```
WITH SAWITH0 AS
  (SELECT T59.TIME_ID AS c2,
          T59.CALENDAR_MONTH_DESC AS c3,
          ROW_NUMBER() OVER (PARTITION BY T59.CALENDAR_MONTH_DESC
                             ORDER BY T59.CALENDAR_MONTH_DESC DESC) AS c4
   FROM SH.TIMES T59 /* Dim - Times */),
     SAWITH1 AS
  (SELECT CASE
              WHEN CASE D1.c4 WHEN 1 THEN D1.c2 ELSE NULL END IS NOT NULL
              THEN Rank() OVER (ORDER BY CASE D1.c4 WHEN 1 THEN D1.c2 ELSE NULL END)
          END AS c1,
          D1.c2 AS c2,
          D1.c3 AS c3
   FROM SAWITH0 D1),
     SAWITH2 AS
  (SELECT min(D1.c1) OVER (PARTITION BY D1.c3) AS c1,
          D1.c2 AS c2
   FROM SAWITH1 D1),
     SAWITH3 AS
  (SELECT D1.c1 + 1 AS c1,
          D1.c2 AS c2
   FROM SAWITH2 D1),
     SAWITH4 AS
  (SELECT T59.CALENDAR_YEAR AS c2,
          T59.CALENDAR_MONTH_DESC AS c3,
          T59.TIME_ID AS c4,
          ROW_NUMBER() OVER (PARTITION BY T59.CALENDAR_MONTH_DESC
                             ORDER BY T59.CALENDAR_MONTH_DESC DESC) AS c5
   FROM SH.TIMES T59 /* Dim - Times */),
     SAWITH5 AS
  (SELECT CASE
              WHEN CASE D1.c5 WHEN 1 THEN D1.c4 ELSE NULL END IS NOT NULL
              THEN Rank() OVER (ORDER BY CASE D1.c5 WHEN 1 THEN D1.c4 ELSE NULL END)
          END AS c1,
          D1.c2 AS c2,
          D1.c3 AS c3
   FROM SAWITH4 D1),
     SAWITH6 AS
  (SELECT DISTINCT min(D1.c1) OVER (PARTITION BY D1.c3) AS c1,
                   D1.c2 AS c2,
                   D1.c3 AS c3
   FROM SAWITH5 D1),
 SAWITH7 AS
  (SELECT sum(T62.AMOUNT_SOLD) AS c1,
          D3.c3 AS c2,
          D3.c2 AS c3
   FROM SH.SALES T62 /* Fact - Sales */,
        SAWITH3 D4,
        SAWITH6 D3
   WHERE (T62.TIME_ID = D4.c2
          AND D3.c1 = D4.c1)
   GROUP BY D3.c2,
            D3.c3),
 SAWITH8 AS
  (SELECT sum(T62.AMOUNT_SOLD) AS c1,
          T59.CALENDAR_MONTH_DESC AS c2,
          T59.CALENDAR_YEAR AS c3
   FROM SH.TIMES T59 /* Dim - Times */,
        SH.SALES T62 /* Fact - Sales */
   WHERE (T59.TIME_ID = T62.TIME_ID)
   GROUP BY T59.CALENDAR_YEAR,
            T59.CALENDAR_MONTH_DESC)SELECT D1.c1 AS c1,
       D1.c2 AS c2,
       D1.c3 AS c3,
       D1.c4 AS c4,
       D1.c5 AS c5
FROM
  (SELECT D1.c1 AS c1,
          D1.c2 AS c2,
          D1.c3 AS c3,
          D1.c4 AS c4,
          D1.c5 AS c5
   FROM
     (SELECT 0 AS c1,
             coalesce(D1.c2, D2.c2) AS c2,
             coalesce(D1.c3, D2.c3) AS c3,
             D2.c1 AS c4,
             D1.c1 AS c5,
             ROW_NUMBER() OVER (PARTITION BY coalesce(D1.c2, D2.c2), coalesce(D1.c3, D2.c3)
                                ORDER BY coalesce(D1.c2, D2.c2) ASC, coalesce(D1.c3, D2.c3)
ASC) AS c6
      FROM SAWITH7 D1
      FULL OUTER JOIN SAWITH8 D2 ON D1.c2 = D2.c2
      AND D1.c3 = D2.c3) D1
   WHERE (D1.c6 = 1)
   ORDER BY c3,
            c2) D1
WHERE rownum <= 65001
```

Self-Service downsides: users write their own analyses

Just one possible way in SQL, there are many others...

```
WITH total_sales AS (
SELECT
  t.calendar_month_desc,
  SUM(s.amount_sold) AS amount_sold
FROM sales s
JOIN times t
  ON s.time_id = t.time_id
GROUP BY t.calendar_month_desc
)
SELECT
  calendar_month_desc,
  amount_sold,
  LAG(amount_sold, 1) OVER (ORDER BY calendar_month_desc) AS amount_sold_mago
FROM total_sales;
```

Self-Service downsides: users write their own analyses

Looking at the explain plan, the cost is from simple to double between the query generated by the self-service analysis versus a manual query.

And this is just a super simple example, it can be a lot worse!

Guest speaker / Secret witness

DIY Analytics:
Control? What's that?

Oracle Analytics

Converged Analytics Platform for all Personas, Workloads and Data

Data Engineers

Business Analysts

Business Users

Governed Analytics		LOB/Self-Service Analytics	
Dashboards	Distributed Pixel-Perfect Reports	Data Visualization	Self-Service Data Preparation
Semantic Models	Query Federation	Storytelling	Direct Connectivity
Briefing Books	Data Export	Collaboration	Mobile

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

Full freedom to users

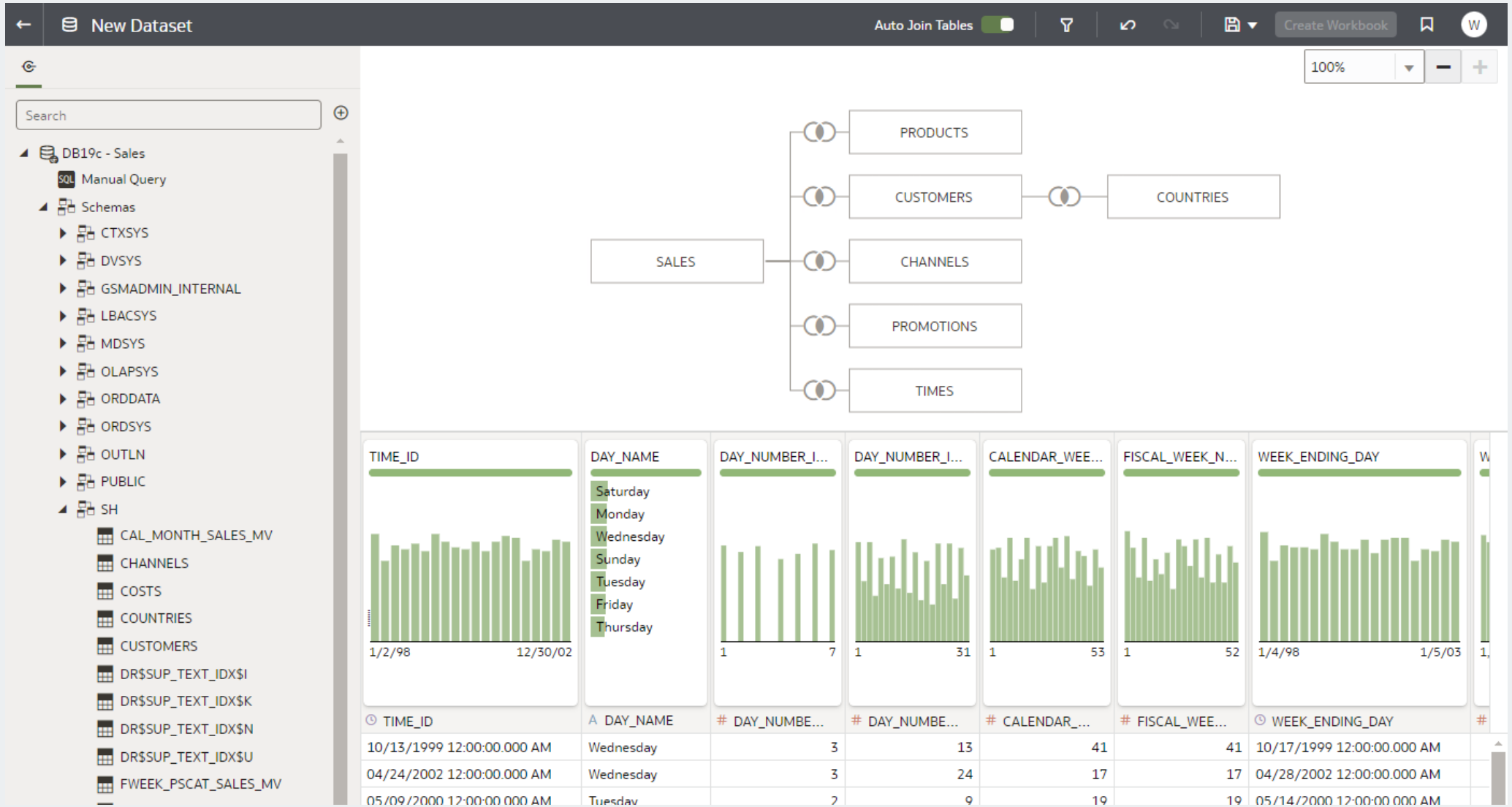
Do It Yourself Analytics isn't an "official" name (as far as I know), but it's what describe this situation.

Users not only can build custom analyses on top of some governed (centrally owned and maintained) metadata.

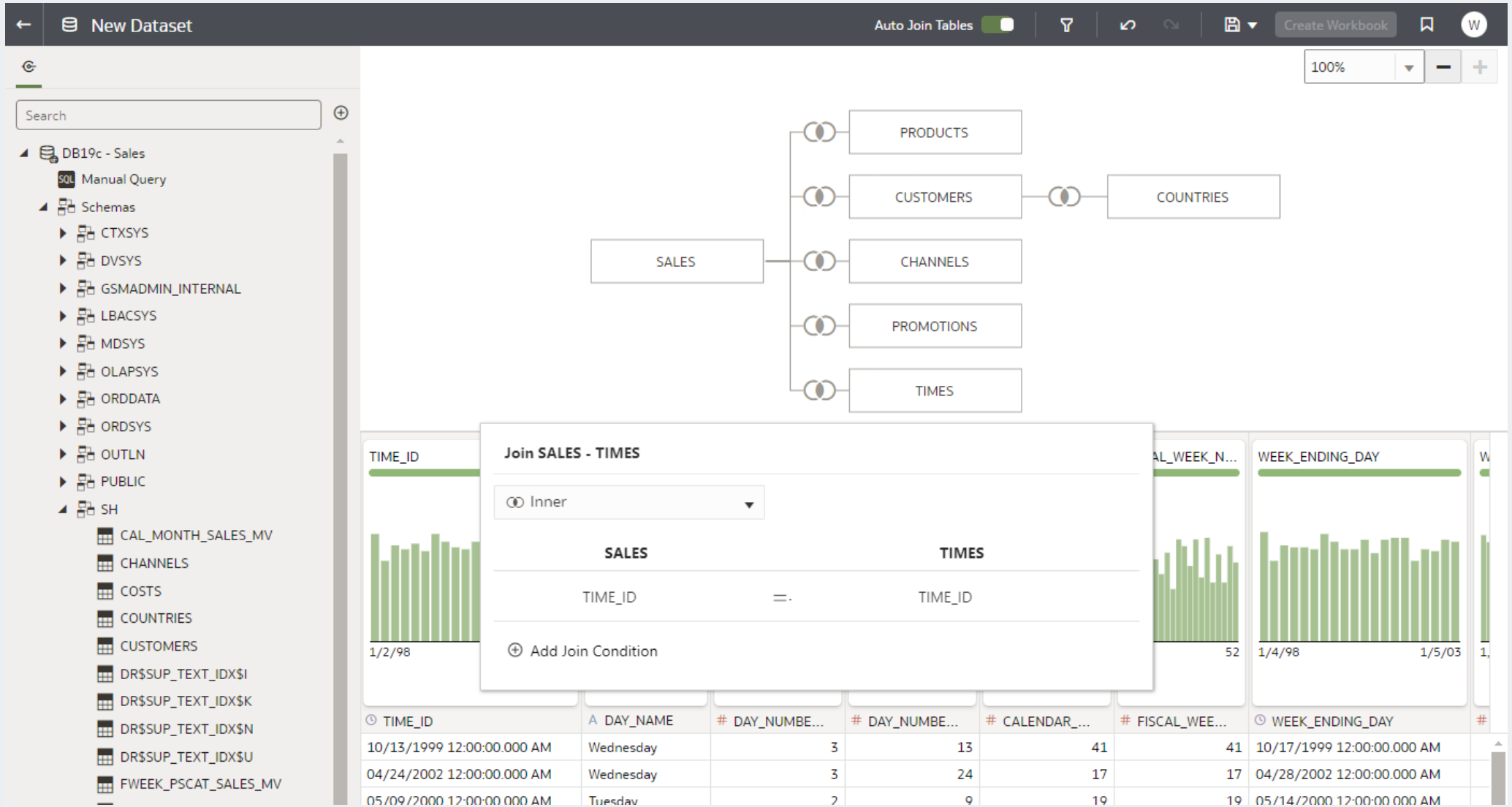
They can bring their own data in the tool:

- Upload Excel spreadsheets
- Connect to a large number of sources
- Define their own data models (connecting various sources)
- Define their own data transformations (ETL)

Data modelling for everybody



Data modelling for everybody



Data modelling for everybody

If your database has PK-FK defined, the tool will automatically connect tables with that.

Nothing prevent a user to build its own join conditions, based on whatever they believe is correct or useful.

Why would a random user use a technical field (primary key in the database) instead of a business field for a join?

Education can help, but you can educate your users as much as you want, in the end, if they believe they know better, they do whatever they want. And what they believe will give them the answer they look for.

ETL for everybody

Lightweight ETL, aka the best way to generate rubbish database tables.

No primary-foreign key, no indexes.

Look at the database audit logs to show the behaviour:

- create insert a temporary table
- drop the destination one
- rename the temporary to the final name

DIY ETL – Indexes, constraints, all pointless things

Worksheet | Query Builder

```
1 select * from user_cons_columns
2 where table_name = 'MY_SALES';
3
```

Query Result

SQL | All Rows Fetched: 12 in 0,078 seconds

	OWNER	CONSTRAINT_NAME	TABLE_NAME	COLUMN_NAME	POSITION
1	SH	SYS_C00171530	MY_SALES	PROD_ID	(null)
2	SH	SYS_C00171531	MY_SALES	CUST_ID	(null)
3	SH	SYS_C00171532	MY_SALES	TIME_ID	(null)
4	SH	SYS_C00171533	MY_SALES	CHANNEL_ID	(null)
5	SH	SYS_C00171534	MY_SALES	PROMO_ID	(null)
6	SH	SYS_C00171535	MY_SALES	QUANTITY_SOLD	(null)
7	SH	SYS_C00171536	MY_SALES	AMOUNT_SOLD	(null)
8	SH	MY_SALES_PROMO_FK	MY_SALES	PROMO_ID	1
9	SH	MY_SALES_CUSTOMER_FK	MY_SALES	CUST_ID	1
10	SH	MY_SALES_PRODUCT_FK	MY_SALES	PROD_ID	1
11	SH	MY_SALES_CHANNEL_FK	MY_SALES	CHANNEL_ID	1
12	SH	MY_SALES_TIME_FK	MY_SALES	TIME_ID	1

DIY ETL – Indexes, constraints, all pointless things

New Data Flow

Show labels 100%

Search

- Add Data
- Join
- Union Rows
- Filter
- Aggregate
- Save Dataset
- Add Columns
- Select Columns
- Rename Columns
- Transform Column
- Merge Columns
- Split Columns
- Bin
- Group
- Branch
- Cumulative Value
- Time Series Forecast
- Analyze Sentiment
- Train Numeric Prediction
- Train Multi-Classifier
- Train Clustering
- Train Binary Classifier
- Apply Model
- Apply AI Model
- Apply Custom Script
- Graph Analytics
- Database Analytics

Sales - SR... → Save Data

Save Dataset

Dataset: Sales - DST

Dataset Table: MY_SALES

Description:

Save data to: Database Connection

Connec...: DB19c - Sales

Table: MY_SALES

When run: Replace existing data

When Run

Columns

Name	Database Name	Treat As	Default Aggregation
PROD_ID	PROD_ID	Attribute	
CUST_ID	CUST_ID	Attribute	
TIME_ID	TIME_ID	Attribute	
CHANNEL_ID	CHANNEL_ID	Attribute	
PROMO_ID	PROMO_ID	Attribute	
QUANTITY_SOLD	QUANTITY_SOLD	Measure	Sum
AMOUNT_SOLD	AMOUNT_SOLD	Measure	Sum

DIY ETL – Indexes, constraints, all pointless things

Where are the foreign keys constraints? Gone…

Worksheet | Query Builder

```
select * from user_cons_columns
where table_name = 'MY_SALES';
```

Query Result

SQL | All Rows Fetched: 7 in 0,022 seconds

	OWNER	CONSTRAINT_NAME	TABLE_NAME	COLUMN_NAME	POSITION
1	SH	SYS_C00171542	MY_SALES	PROD_ID	(null)
2	SH	SYS_C00171543	MY_SALES	CUST_ID	(null)
3	SH	SYS_C00171544	MY_SALES	TIME_ID	(null)
4	SH	SYS_C00171545	MY_SALES	CHANNEL_ID	(null)
5	SH	SYS_C00171546	MY_SALES	PROMO_ID	(null)
6	SH	SYS_C00171547	MY_SALES	QUANTITY_SOLD	(null)
7	SH	SYS_C00171548	MY_SALES	AMOUNT_SOLD	(null)

DIY ETL – Indexes, constraints, all pointless things

- Create a new table with a temporary name and load the data
- If the target table exists, drop it
- Rename the table created to the final name

Worksheet | Query Builder

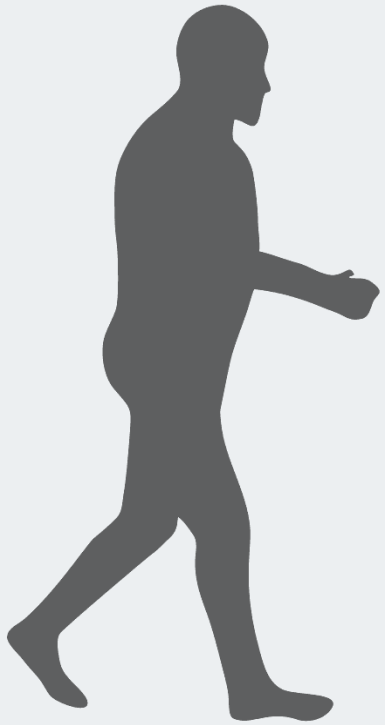
```
select * from user_cons_columns
where table_name = 'MY_SALES';
```

Query Result

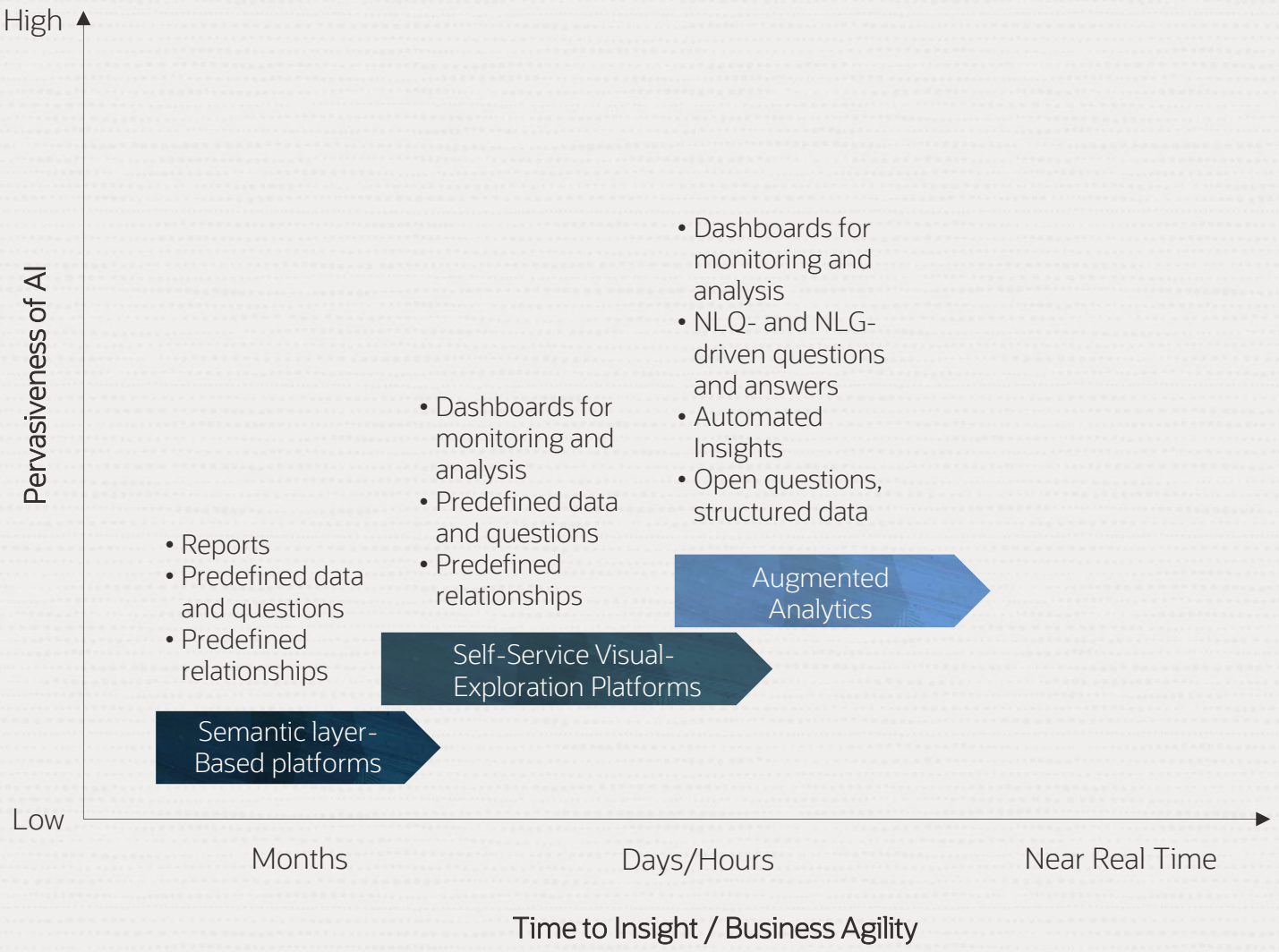
SQL | All Rows Fetched: 7 in 0,022 seconds

	OWNER	CONSTRAINT_NAME	TABLE_NAME	COLUMN_NAME	POSITION
1	SH	SYS_C00171542	MY_SALES	PROD_ID	(null)
2	SH	SYS_C00171543	MY_SALES	CUST_ID	(null)
3	SH	SYS_C00171544	MY_SALES	TIME_ID	(null)
4	SH	SYS_C00171545	MY_SALES	CHANNEL_ID	(null)
5	SH	SYS_C00171546	MY_SALES	PROMO_ID	(null)
6	SH	SYS_C00171547	MY_SALES	QUANTITY_SOLD	(null)
7	SH	SYS_C00171548	MY_SALES	AMOUNT_SOLD	(null)

Augmented Analytics:
No ML? No fun!



Timeline of Innovation Points in the Analytics Market



Source: Gartner 2021

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

Where are the skills ?

Only few resources hold technical skills to execute ML / AI / Advanced Analytics…

… yet many Business-Analysts can derive great value of using AI/ML for their domains

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

Oracle Analytics

Converged Analytics Platform for all Personas, Workloads and Data

Data Engineers		Business Analysts		Business Users	Citizen Data Scientists	
Governed Analytics		**LOB/Self-Service Analytics**			**Augmented Analytics**	
Dashboards	Distributed Pixel-Perfect Reports	Data Visualization	Self-Service Data Preparation		Voice & Chatbot	Natural Language
Semantic Models	Query Federation	Storytelling	Direct Connectivity		Data Profiling & Enrichment	AI Explainability
Briefing Books	Data Export	Collaboration	Mobile		Automated Insights	Machine Learning

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

One-Click ML

BC_ML

Search

- Add Data
- Join
- Union Rows
- Filter
- Aggregate
- Save Dataset
- Create Essbase Cube
- Add Columns
- Select Columns
- Rename Columns
- Transform Column
- Merge Columns
- Split Columns
- Bin
- Group
- Branch
- Cumulative Value
- Time Series Forecast
- Analyze Sentiment
- Train Numeric Prediction
- Train Multi-Classifier
- Train Clustering
- Train Binary Classifier

Ex2Data1 → Train Bina... Classifier → Save Model

Train Binary Classifier

Model Training Script: Logistic Regression for model training

* Target: res

target, the target(label) to learn/predict

Positive Class in Target: Yes

Positive class in the target value. Default is Yes.

Predict Value Threshold %: 50

The threshold value to determine the predict values

Maximum Null Value Percent: 50

Maximum number of records in percent that can contain null values.

Numerical Column Imputation: Mean

The mode method for numeric features to fill NA. Four options: mean, max, min, median. Default is mean.

Categorical Encoding Method: Indexer

Encoding method.

Categorical Column Imputation: Most Frequent

One-Click ML

Auto ML predicted affinity card Data Flow

Search

- Add Data
- Join
- Union Rows
- Filter
- Aggregate
- Save Dataset
- Create Essbase Cube
- Add Columns
- Select Columns
- Rename Columns
- Transform Column
- Merge Columns
- Split Columns
- Bin
- Group
- Branch
- Cumulative Value
- Time Series Forecast
- Analyze Sentiment
- Train Numeric Prediction
- Train Multi-Classifier
- Train Clustering
- Train Binary Classifier
- AutoML
- Apply Model

ML test da... → Apply Model → Save Data

Apply Model

Model Affinity Card ML Model

Outputs

Create	Output	Column Name
☑	Prediction	Prediction
☑	PredictionProbability	PredictionProbability

Additional Outputs

Parameters

Cost Model - Auto: On

Use the option Cost Model Auto

Compute lift and gain: No

Use this option to generate model lift and gain values for this dataset. Note:- An additional output dataset with the same name and suffix _LIFT will be created.

Target column to compute lift: Select a column

Column containing actual values to be used to compute lift. This info is required to compute lift.

One-Click ML

ML is just one click away.

Where does it run? What does it do? Who cares... Not a user problem!

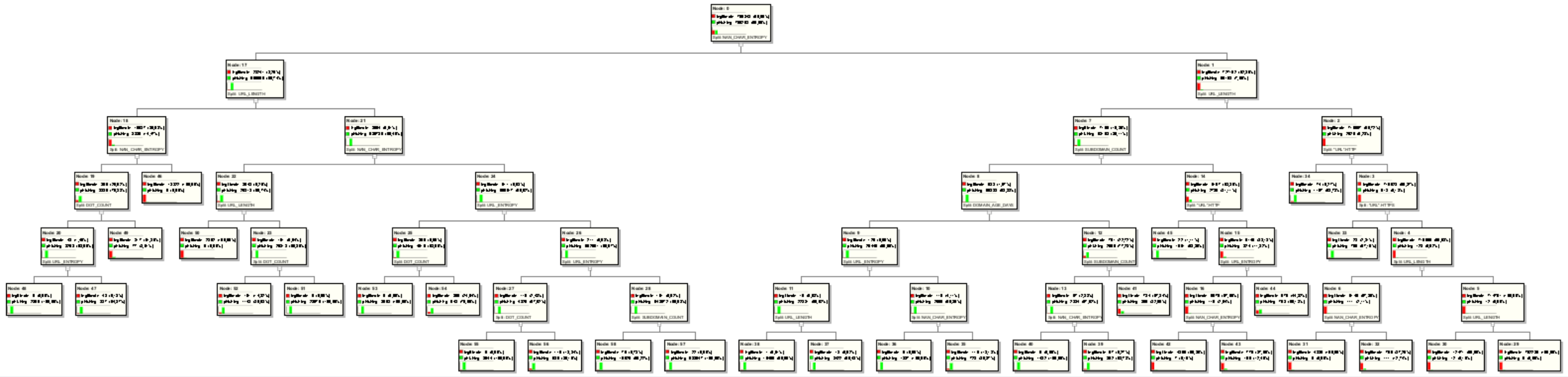
Brand new teams of data scientists have been created.

They ingest data at a crazy pace, they often don't care if the database can do the job, they will export you whole database into their python playground and build their ML models there.

- Or maybe they will decide to train a ML model on your database, on all the data
- All the records will be retrieved and a lot of calculation will happen
- Even the simplest decision tree can be a quite long list of IF ... ELSE ...

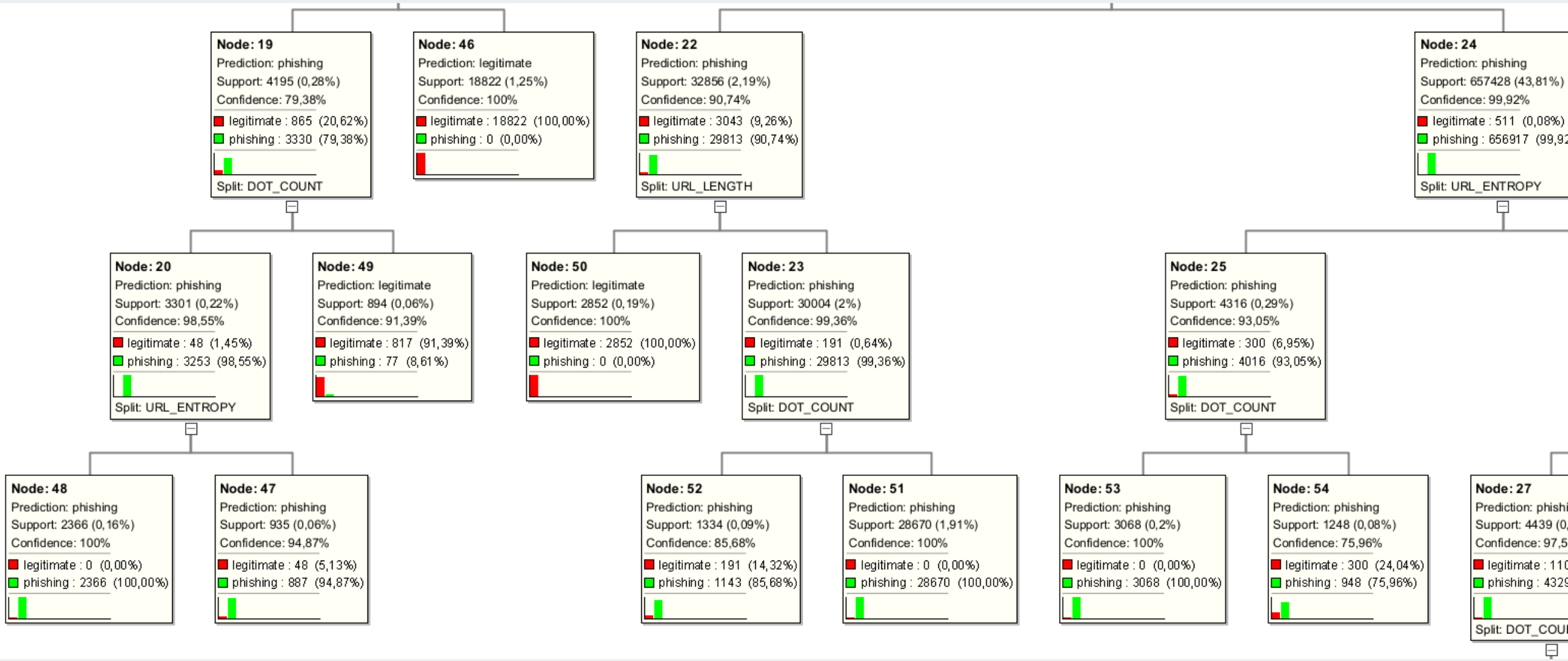
ML: decision tree for classification

An example of a Decision Tree model



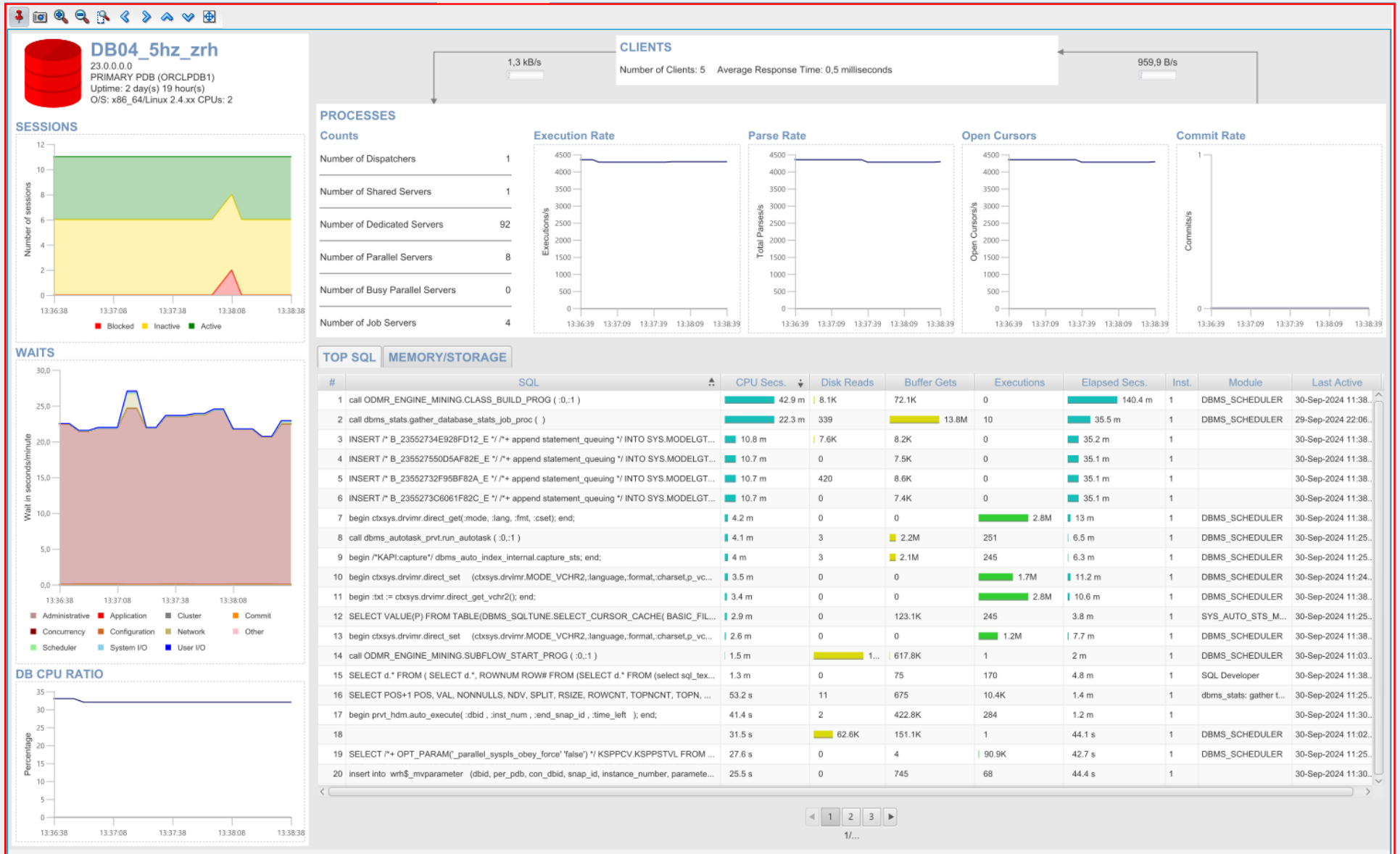
ML: decision tree for classification

An example of a Decision Tree model



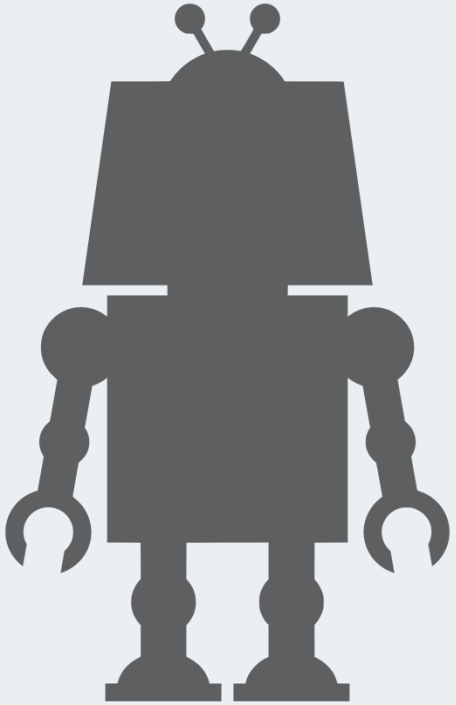
ML: decision tree for classification

I don’t know how to fully read it, but the DB was busy when training the model.

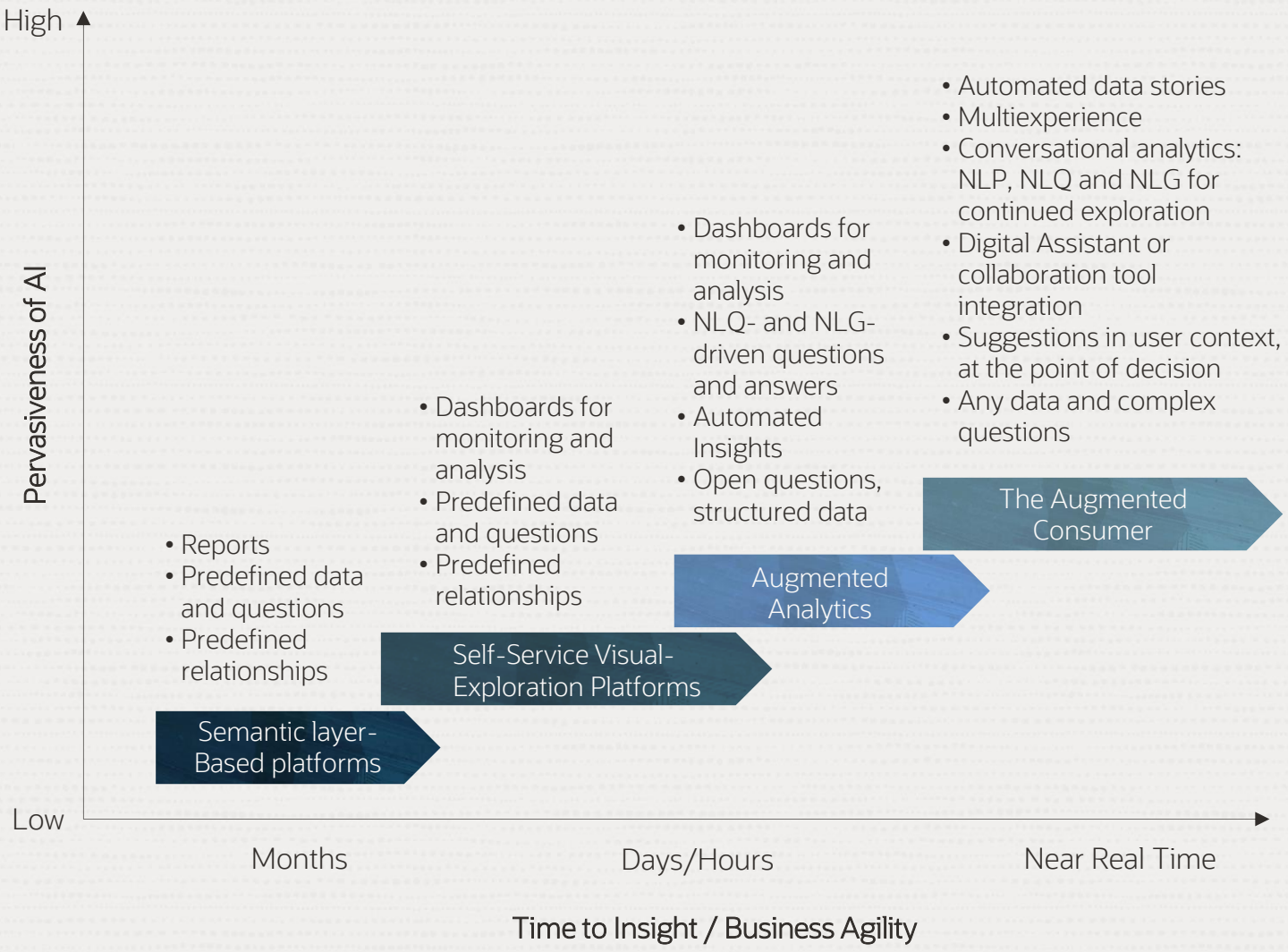


#	SQL	CPU Secs.	Disk Reads	Buffer Gets	Executions	Elapsed Secs.	Inst.	Module	Last Active
1	call ODMR_ENGINE_MINING.CLASS_BUILD_PROG (:0,:1)	42.9 m	8.1K	72.1K	0	140.4 m	1	DBMS_SCHEDULER	30-Sep-2024 11:38..
2	call dbms_stats.gather_database_stats_job_proc ( )	22.3 m	339	13.8M	10	35.5 m	1	DBMS_SCHEDULER	29-Sep-2024 22:06..
3	INSERT /* B_23552734E928FD12_E */ /*+ append statement_queuing */ INTO SYS.MODELGT...	10.8 m	7.6K	8.2K	0	35.2 m	1		30-Sep-2024 11:38..
4	INSERT /* B_235527550D5AF82E_E */ /*+ append statement_queuing */ INTO SYS.MODELGT...	10.7 m	0	7.5K	0	35.1 m	1		30-Sep-2024 11:38..
5	INSERT /* B_23552732F95BF82A_E */ /*+ append statement_queuing */ INTO SYS.MODELGT...	10.7 m	420	8.6K	0	35.1 m	1		30-Sep-2024 11:38..
6	INSERT /* B_2355273C6061F82C_E */ /*+ append statement_queuing */ INTO SYS.MODELGT...	10.7 m	0	7.4K	0	35.1 m	1		30-Sep-2024 11:38..
7	begin ctxsys.drvimr.direct_get(:mode, :lang, :fmt, :cset); end;	4.2 m	0	0	2.8M	13 m	1	DBMS_SCHEDULER	30-Sep-2024 11:38..
8	call dbms_autotask_prvt.run_autotask (:0,:1)	4.1 m	3	2.2M	251	6.5 m	1	DBMS_SCHEDULER	30-Sep-2024 11:25..
9	begin /*KAPI:capture*/ dbms_auto_index_internal.capture_sts; end;	4 m	3	2.1M	245	6.3 m	1	DBMS_SCHEDULER	30-Sep-2024 11:25..
10	begin ctxsys.drvimr.direct_set (ctxsys.drvimr.MODE_VCHR2,:language,:format,:charset,p_vc...	3.5 m	0	0	1.7M	11.2 m	1	DBMS_SCHEDULER	30-Sep-2024 11:24..
11	begin :txt := ctxsys.drvimr.direct_get_vchr2(); end;	3.4 m	0	0	2.8M	10.6 m	1	DBMS_SCHEDULER	30-Sep-2024 11:38..
12	SELECT VALUE(P) FROM TABLE(DBMS_SQLTUNE.SELECT_CURSOR_CACHE(BASIC_FIL...	2.9 m	0	123.1K	245	3.8 m	1	SYS_AUTO_STS_M...	30-Sep-2024 11:25..
13	begin ctxsys.drvimr.direct_set (ctxsys.drvimr.MODE_VCHR2,:language,:format,:charset,p_vc...	2.6 m	0	0	1.2M	7.7 m	1	DBMS_SCHEDULER	30-Sep-2024 11:38..
14	call ODMR_ENGINE_MINING.SUBFLOW_START_PROG (:0,:1)	1.5 m	1...	617.8K	1	2 m	1	DBMS_SCHEDULER	30-Sep-2024 11:03..
15	SELECT d.* FROM (SELECT d.*, ROWNUM ROW# FROM (SELECT d.* FROM (select sql_tex...	1.3 m	0	75	170	4.8 m	1	SQL Developer	30-Sep-2024 11:38..
16	SELECT POS+1 POS, VAL, NONNULLS, NDV, SPLIT, RSIZE, ROWCNT, TOPNCNT, TOPN, ...	53.2 s	11	675	10.4K	1.4 m	1	dbms_stats: gather t...	30-Sep-2024 11:25..
17	begin prvt_hdm.auto_execute(:dbid , :inst_num , :end_snap_id , :time_left); end;	41.4 s	2	422.8K	284	1.2 m	1		30-Sep-2024 11:30..
18		31.5 s	62.6K	151.1K	1	44.1 s	1	DBMS_SCHEDULER	30-Sep-2024 11:02..
19	SELECT /*+ OPT_PARAM('_parallel_syspls_obey_force' 'false') */ KSPPCV.KSPPSTVL FROM ...	27.6 s	0	4	90.9K	42.7 s	1	DBMS_SCHEDULER	30-Sep-2024 11:25..
20	insert into wrh$_mvparameter (dbid, per_pdb, con_dbid, snap_id, instance_number, paramete...	25.5 s	0	745	68	44.4 s	1		30-Sep-2024 11:30..

AI:
If it doesn’t have vectors, it can’t be right!



Timeline of Innovation Points in the Analytics Market



Source: Gartner 2021

Slide author: Philippe Lions, Vice President, Oracle Analytics PM

How big is your storage?

Vectors, vectors everywhere, vectors for everything

- OpenAI `text-embedding-3-large` return a vector of dimensionality 3072
- 'POUG2024', 8 characters in, return a vector that when converted to CLOB is 50'335 characters long!
- `cohere.embed-english-v3.0` returns vectors with 1024 dimensions, only 16'808 characters…

Worksheet | Query Builder

```
112 WITH t AS (
113   SELECT
114     FROM_VECTOR(
115       DBMS_VECTOR.UTL_TO_EMBEDDING('POUG2024',
116         json('{
117                "provider": "OpenAI",
118                "credential_name": "CRED_OPENAI_PRJ",
119                "url": "https://api.openai.com/v1/embeddings",
120                "model": "text-embedding-3-large"
121               }')
122       ) RETURNING CLOB) as v
123 )
124 SELECT DBMS_LOB.GETLENGTH(v), v FROM t;
125
```

Query Result

SQL | All Rows Fetched: 1 in 1,179 seconds

	DBMS_LOB.GETLENGTH(V)	V
1	50335	[-4.13695946E-002,1.78114194E-002,-1.69239249E-002,1.66954603E-002,4.654...

How big is your storage?

There are many other models generating embeddings of various sizes.

`all_MiniLM_L12_v2` can be loaded in the database and executed there.

Return vectors of dimensionality 384, such a vector in a textual form would be ~6’292 chars.

For an input of 256 tokens (~words) maximum, a vector represented by 6’292 chars is produced.

Indexing vectors? Ok, but which distance metric?

Having millions or more vectors is good, but could become slow for proximity queries.

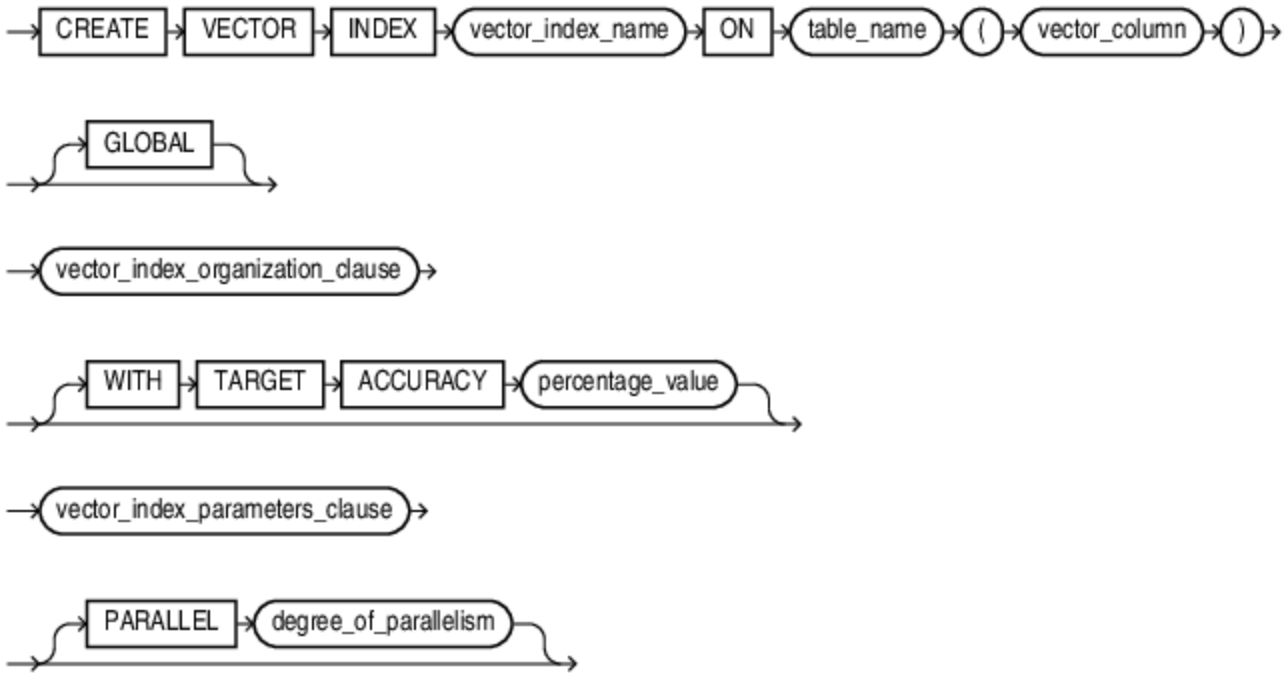
Vector indexes speed up vector search.

But ...
You maybe will need to define multiple indexes on the same vector column of the same table, because vector indexes are “specialised” indexes and not generic.

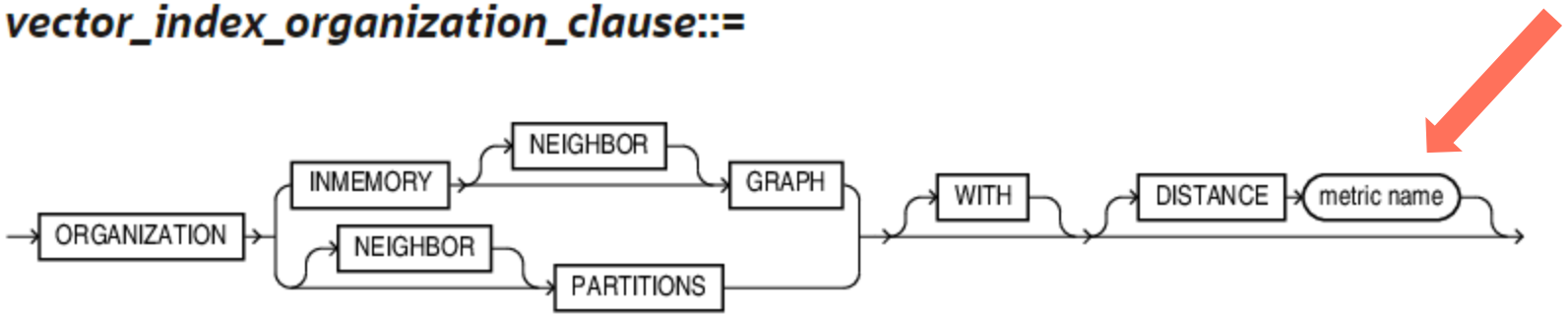
The metric is the key.

Indexing vectors? Ok, but which distance metric?

Syntax



vector_index_organization_clause::=



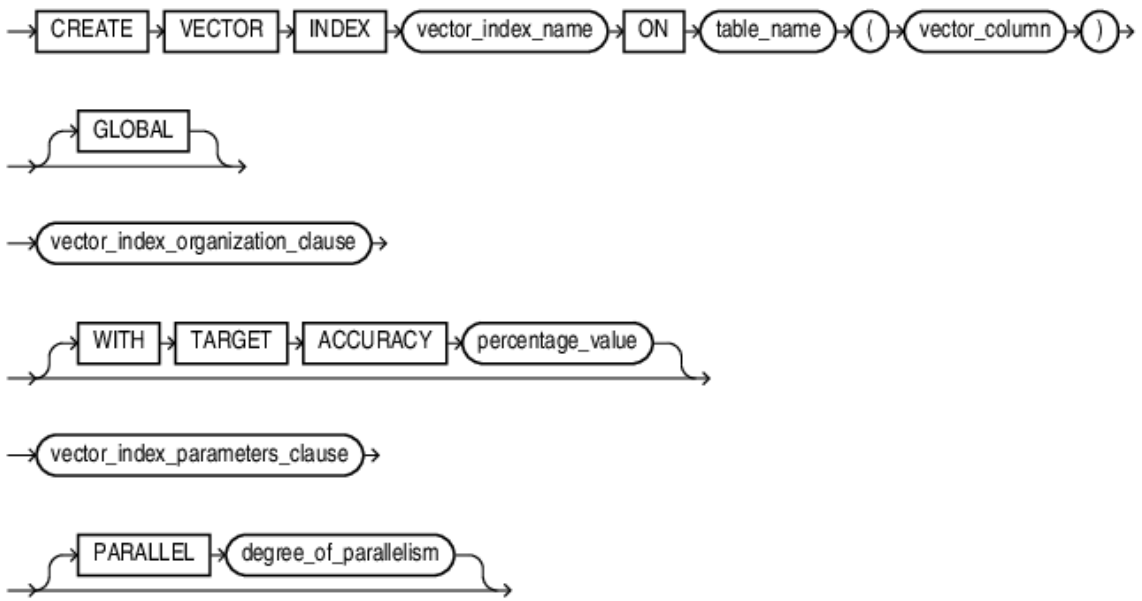
Indexing vectors? Ok, but which distance metric?

A vector index is defined saying which distance metric it should use.

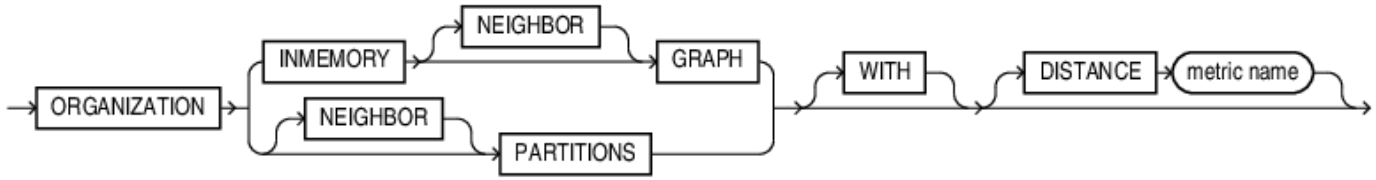
- Euclidian
- Euclidian squared
- Cosine
- Dot Product
- Manhattan

An index with a mismatching metric will not be used.

Syntax



vector_index_organization_clause::=



AI in the database leads to XY problems

AI is often performed by 3rd party (web)services.

Many applications ask for an API key for service XYZ to be able to provide some AI functionalities. Inside the Oracle Database it isn't different.

This often leads to XY problems between users and DBAs.

What is the XY problem?

- The XY problem is asking about your attempted solution (Y) rather than your actual problem (X). This leads to enormous amounts of wasted time and energy, both on the part of people asking for help, and on the part of those providing help.

Security driven by users

A simple example...

Oracle Database has DBMS_VECTOR and DBMS_VECTOR_CHAINS able to perform embedding and LLM-based text generation calling external webservices.

The documentation doesn't explain in detail what "good practice" security must be configured to allow the calls.

But, luckily a number of blog posts and Oracle resources provide a solution!

Security driven by us

oracle-japan.github.io/ocitutorials/ai-vector-search

ORACLE DATABA
VECTOR SEARCH

101 : Always Free
ADBインスタンス
てみよう

102 : 仮想マシン
Database 23ai Fre
トールしてみよう

103 : Oracle AI Ve
の基本操作を試し

**104 :ファイル→テ
ャンク→ベクトル
よびベクトル検索**

105: マルチベクト
複数のドキュメン
してみよう

106 : Oracle Data
LangChainでRAG
てみよう

107 : 会話履歴保持
を取り入れたRAG
してみよう

108 : SELECT AI w
試してみよう

qiita.com/kenwatan/items/d1bc626ab18628c84958

事前作業

R_ROLE」と

imited on USEI

DBMS_VECTOR_CHAIN.UTL_TO_SUMMARY
実行手順（OCI GenAIの場合）

1. DBMS_NETWORK_ACL_ADMINを使用してホストに権限付与

```
BEGIN DBMS_NETWORK_ACL_ADMIN.APPEND_HOST_ACE( host => '*', ace => xs$ace_type(privilege_list =>
xs$name_list('connect'), principal_name => 'docuser', principal_type => xs_acl.ptype_db)); END; /
```

2. DBMS_VECTOR_CHAIN.CREATE_CREDENTIALを使用してOCIの資格証明を作成

```
declare
  jo json_object_t;
begin
  -- create an OCI credential
  jo := json_object_t();
  jo.put('user_ocid','ocid1.user.oc1..aabbalbbaa1112233aabbaabb1111222aa1111bb');
  jo.put('tenancy_ocid','ocid1.tenancy.oc1..aaaaalbbbb1112233aaaabbaa1111222aaa111a');
  jo.put('compartment_ocid','ocid1.compartment.oc1..ababalabab1112233abababab1111222aba11ab');
  jo.put('private_key','AAAaaaBBB11112222333...AAA111AAABBB222aaa1a/+');
  jo.put('fingerprint','01:1a:a1:aa:12:a1:12:1a:ab:12:01:ab:a1:12:ab:1a');
  dbms_output.put_line(jo.to_string);
  dbms_vector_chain.create_credential(
    credential_name   => 'OCI_CRED',
    params            => json(jo.to_string));
end;
/
```

61

Security driven by users

1. Grant the `CREATE CREDENTIAL` privilege; and

```
grant create credential to &app_schema.;
```

2. Allow the schema to make network connections using the `DBMS_NETWORK_ACL_ADMIN` procedure.

```
begin
    dbms_network_acl_admin.append_host_ace(
        host => '*'
        , ace => xs$ace_type(
            privilege_list => xs$name_list('connect')
            , principal_name => '&app_schema.'
            , principal_type => xs_acl.ptype_db
        )
    );
end;
/
```

There were no issues performing the second task, however, for the first, the `ADMIN` user itself does not have that privilege. Hence, when you execute the procedure `DBMS_VECTOR_CREATE_CREDENTIAL` you would have received this:

Security driven by users

Apparently, the solution to be able to use a 3rd party webservice for AI, is to allow your schema to connect to any random resource using a list of packages:

- UTL_TCP
- UTL_SMTP
- UTL_MAIL
- UTL_HTTP

Do you really need all that?

No! But the user will ask the DBA to execute that statement!

- A XY problem

1. Grant the `CREATE CREDENTIAL` privilege; and

```
grant create credential to &app_schema.;
```

2. Allow the schema to make network connections using the `DBMS_NETWORK_ACL_ADMIN` procedure.

```
begin
    dbms_network_acl_admin.append_host_ace(
        host => '*'
        , ace => xs$ace_type(
            privilege_list => xs$name_list('connect')
            , principal_name => '&app_schema.'
            , principal_type => xs_acl.ptype_db
        )
    );
end;
/
```

There were no issues performing the second task, however, for the first, the `ADMIN` user itself does not have that privilege. Hence, when you execute the procedure `DBMS_VECTOR_CREATE_CREDENTIAL` you would have received this:

Security driven by users

Why not just allow the minimal possible opening?

```
BEGIN
  -- for OCI GenAI
  DBMS_NETWORK_ACL_ADMIN.append_host_ace (
    host       => 'inference.generativeai.eu-frankfurt-1.oci.oraclecloud.com',
    lower_port => 443,
    upper_port => 443,
    ace        => xs$ace_type(privilege_list => xs$name_list('http'),
                              principal_name => 'my_user',
                              principal_type => xs_acl.ptype_db));

  -- for Cohere
  DBMS_NETWORK_ACL_ADMIN.append_host_ace (
    host       => 'api.cohere.com',
    lower_port => 443,
    upper_port => 443,
    ace        => xs$ace_type(privilege_list => xs$name_list('http'),
                              principal_name => 'my_user',
                              principal_type => xs_acl.ptype_db));
END;
/
```

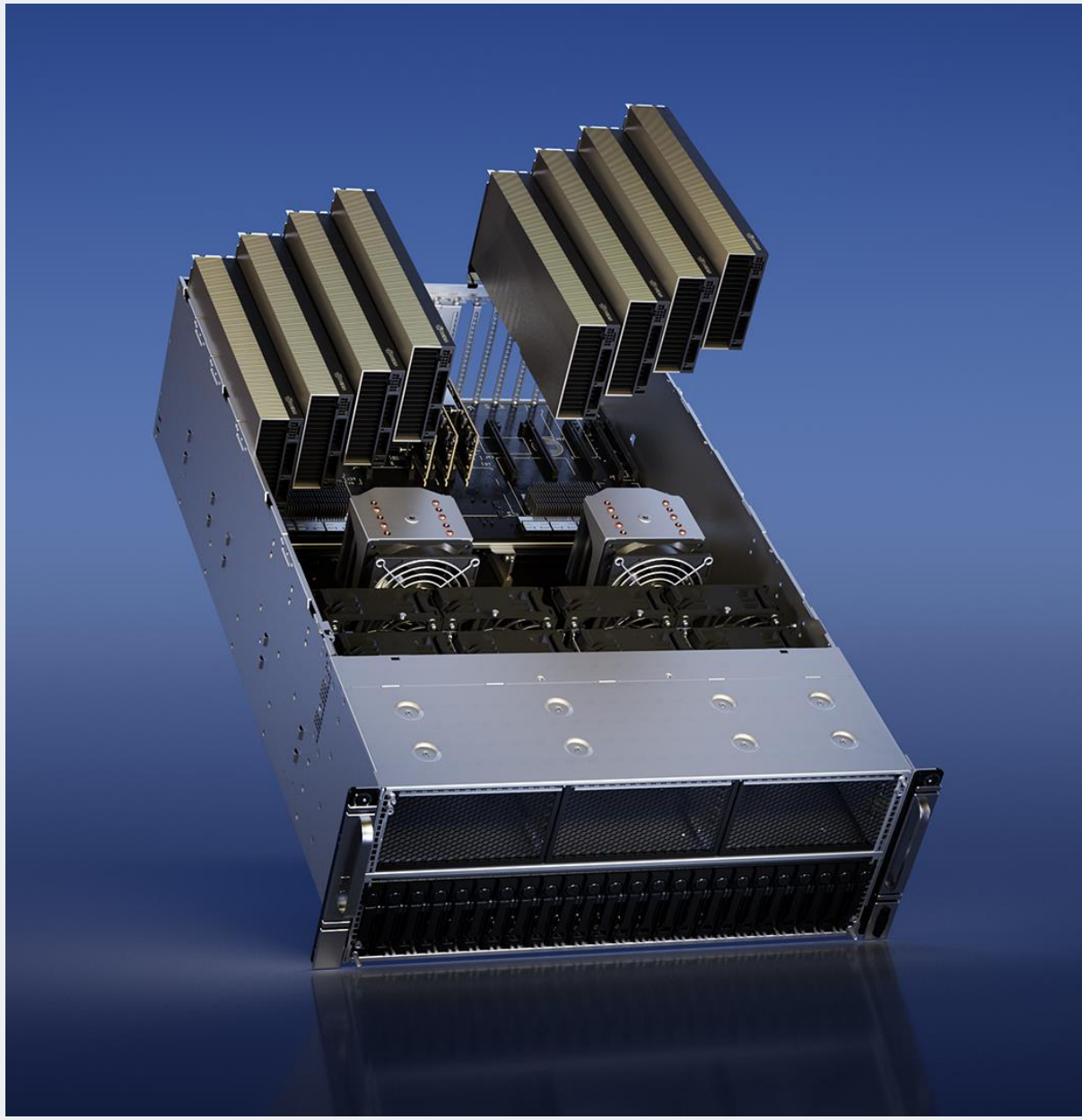
AI has different needs

Does your database server look like this?

Oracle GPU License?

A GPU expansion module for Exadata?

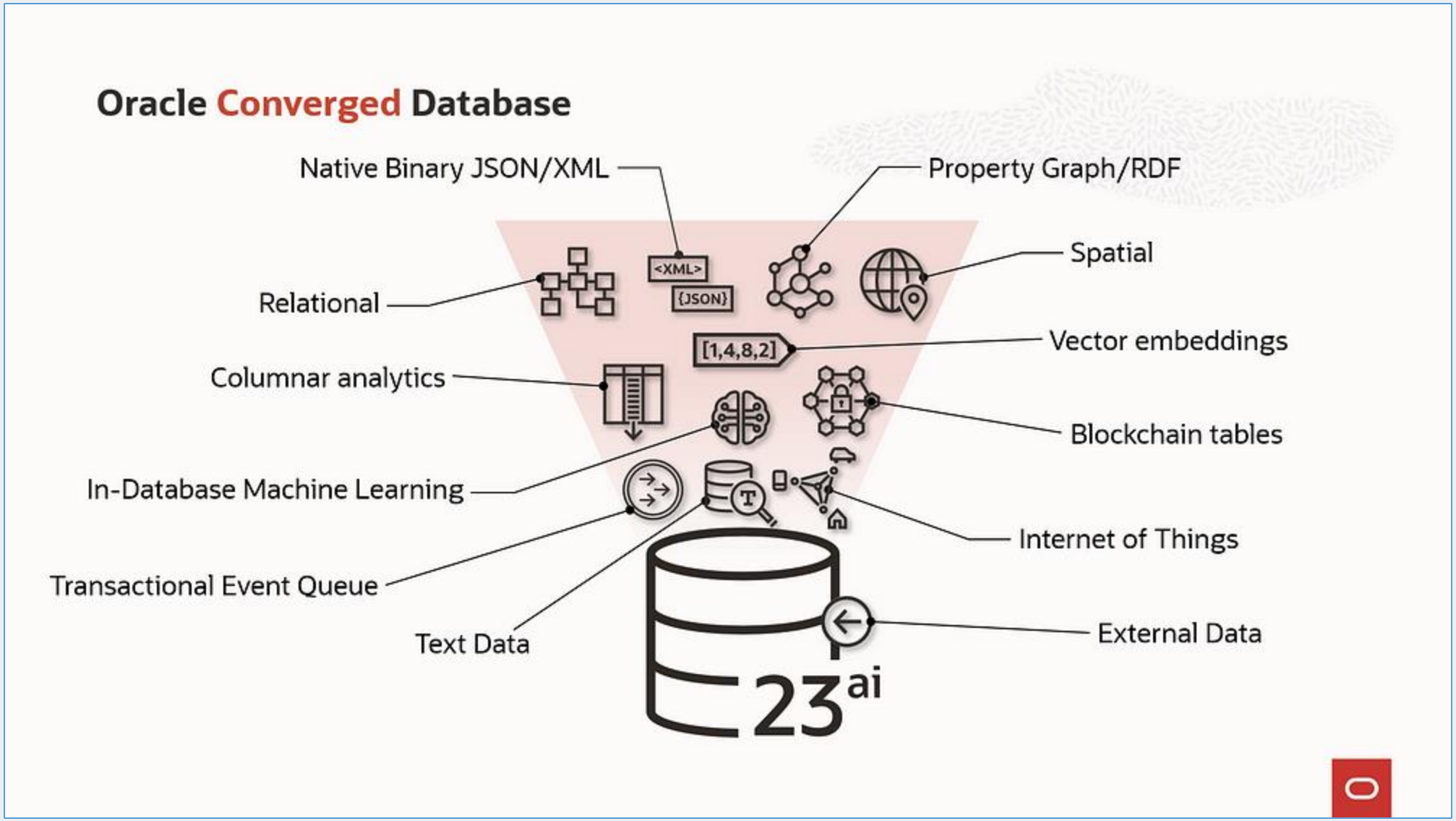
How are you supposed to size your database hardware if your usage is a moving target?



And there is more...

Converged database

Oracle Converged Database



Oracle Converged Database

Oracle Converged Database is great: many workloads in a single product.

But...

It's like a Swiss Army knife: practical and very useful when you have nothing else, but struggle to keep up with tools dedicated to a single task.

Oracle Converged Database

VS

Oracle Converged Database

Oracle Converged Database is great: many workloads in a single product.

But...

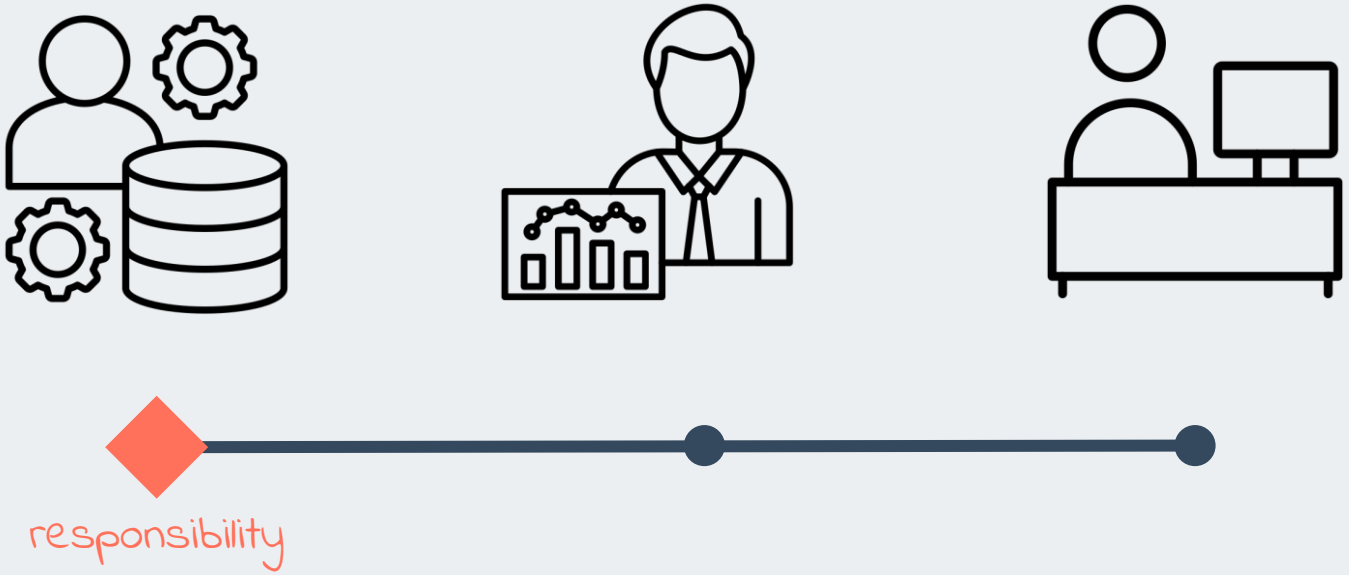
It's like a Swiss Army knife: practical and very useful when you have nothing else, but struggle to keep up with tools dedicated to a task.

Not saying that the converged database is wrong, but if at some point you are using your Oracle Database for everything but a relational database, you maybe bought the wrong product?

Analytics in 2024 can hurt your database...

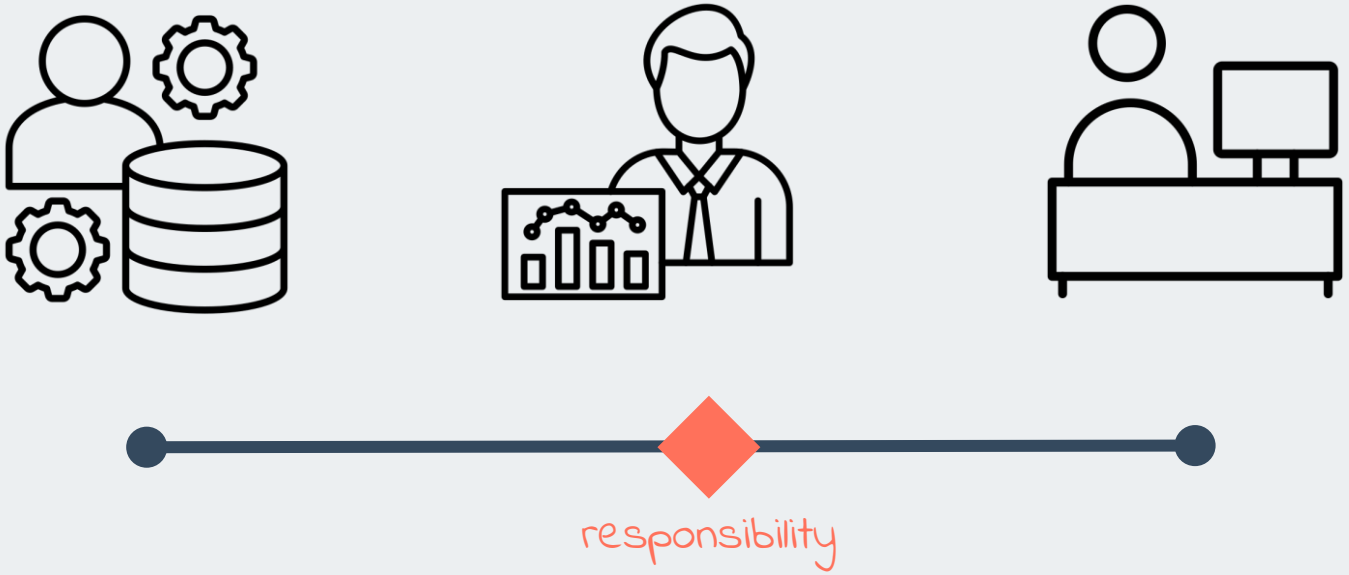
Back to many years ago...

When it was all about Reporting, the DBA had all the information to proactively prepare the database to perform that task. Tuning queries, spreading execution across the available window.



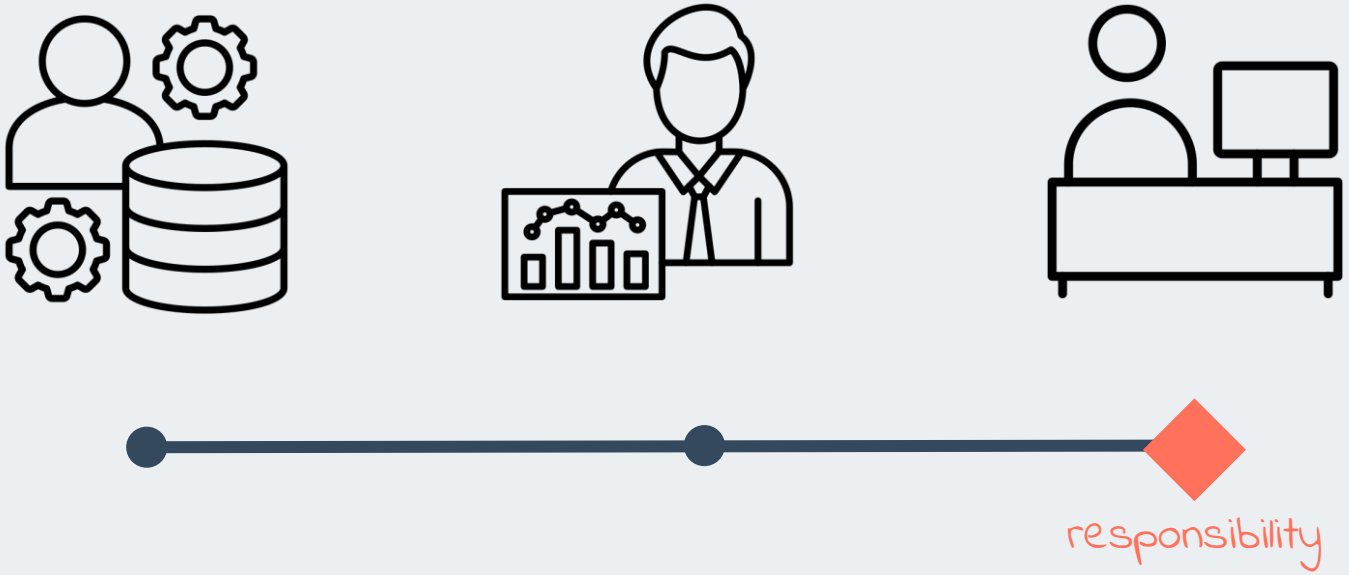
It started shifting to something else...

With Business Intelligence, things started changing. The data expert was the one modelling things, influencing queries. But the DBA could still play a role, observing the workload and suggesting changes and working hand in hand with the data expert to achieve the best results.



To become user-driven...

With Self-service Analytics first, and even worse with DIY Analytics, the end users are free to do whatever they want. Data experts can try to train them, to collect requirements and build content in the “old” way to make it available to them. DBA are reacting more than being proactive.



And finally be out of control!

With ML and AI, who is in charge? Who has the control? Not even the end user does know anymore what's going on. With the introduction of AI Assistants, it's a LLM generating queries! DBAs can only try to run after all that...

Two users needing the same thing could get different queries just because they worded slightly differently the question.