

SQL

Announcements

Databases

Data is very, very powerful!

AI is made of data...

DOI:10.1145/3448247

Industry experiences on the data challenges of AI and the call for a data ecosystem for industrial enterprises.

BY CHRISTOPH GRÖGER

There Is No AI Without Data

ARTIFICIAL INTELLIGENCE (AI) has evolved from hype to reality over the past few years. Algorithmic advances in machine learning and deep learning, significant increases in computing power and storage, and huge amounts of data generated by digital transformation efforts make AI a game-changer across all industries. AI has the potential to radically improve business processes with, for instance, real-time quality prediction in manufacturing, and to enable new business models,

such as connected car services and self-optimizing machines. Traditional industries, such as manufacturing, machine building, and automotive, are facing a fundamental change: from the production phase up to the delivery of AI-enhanced processes and services as part of Industry 4.0.³⁰ This paper focuses on AI for industrial enterprises with a special emphasis on machine learning and data mining.

Despite the great potential of AI and the large investments in AI technologies undertaken by industrial enterprises, AI has not yet delivered on the promises in industry practice. The core business of industrial enterprises is not yet AI-enhanced. AI solutions instead constitute islands for isolated cases—such as automated inspection of selected machines—in the factory with varying success. According to current industry surveys, data issues constitute the main reasons for the insufficient adoption of AI in industrial enterprises.^{27,35}

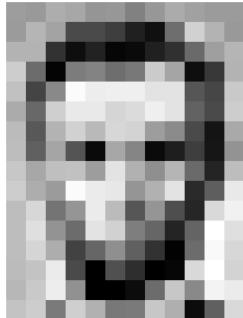
In general, it is nothing new that data preparation and data quality are key for AI and data analytics, as there is no AI without data. This has been an issue since the early days of business intelligence (BI) and data warehousing.³ However, the manifold data challenges of AI in industrial enterprises go far beyond detecting and repairing dirty data. This article profoundly investi-

» key insights

- Despite AI's great potential, the business of industrial enterprises is not yet AI-enhanced. AI is done in an insular fashion, leading to a siloed and heterogeneous enterprise data landscape that limits the comprehensive application of AI.
- Data challenges, such as data management, data democratization, and data governance, are among the major obstacles to leveraging AI and go far beyond ensuring data quality, comprising diverse aspects such as metadata management, data architecture, and data ownership.
- The presented data ecosystem for industrial enterprises addresses these challenges and comprises three components that comprises data producers, data consumers, and data roles for AI.

IMAGE OF ALBERT ANDREWS EDITED WITH ADDITIONAL PHOTOS BY PHOTONSTOCK.COM

98 COMMUNICATIONS OF THE ACM | NOVEMBER 2021 | VOL. 64 | NO. 11



157	153	174	168	150	152	129	151	172	161	155	156
155	182	163	74	75	62	93	17	10	210	180	154
180	180	50	14	34	6	10	33	48	106	159	181
206	109	6	124	131	111	120	204	166	16	66	180
194	68	137	251	237	239	239	228	227	67	71	201
172	105	207	23	239	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	155	54	10	162	154	11	31	62	22	148
199	168	191	195	158	227	178	143	182	105	36	190
205	174	155	252	236	231	149	178	228	43	45	234
190	216	116	141	236	187	85	150	79	38	210	241
190	224	147	150	227	210	127	102	36	101	255	224
190	214	173	56	103	143	96	50	2	109	249	215
187	196	235	73	1	81	47	0	6	217	255	211
183	202	237	143	0	9	72	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

157	153	174	168	150	152	129	151	172	161	155	156
180	180	50	14	34	6	10	33	48	106	159	181
206	109	5	124	131	111	120	204	166	16	66	180
206	109	5	124	131	111	120	204	166	16	66	180
194	68	137	251	237	239	239	228	227	67	71	201
172	105	207	23	239	214	220	239	228	98	74	206
188	88	179	209	185	215	211	158	139	75	20	169
189	97	155	54	10	162	154	11	31	62	22	148
199	168	191	195	158	227	178	143	182	105	36	190
205	174	155	252	236	231	149	178	228	43	45	234
190	216	116	141	236	187	85	150	79	38	218	241
190	224	147	150	227	210	127	102	36	101	255	224
190	214	173	56	103	143	96	50	2	109	249	215
187	196	235	73	1	81	47	0	6	217	255	211
183	202	237	143	0	9	72	108	200	138	243	236
195	206	123	207	177	121	123	200	175	13	96	218

digital images are made out of data...

To many of the biggest, most powerful corporations in the world...



...data about *us* is their most prized resource!

Database Management Systems

Database management systems (DBMS) are important, heavily used, and interesting!

A table is a collection of records, which are rows that have a value for each column

Latitude	Longitude	Name
38	122	Berkeley
42	71	Cambridge
45	93	Minneapolis

The Structured Query Language (SQL) is perhaps the most widely used programming language

SQL is a *declarative* programming language

Declarative Programming

In **declarative programming**:

- A "program" is a description of the desired result
- The interpreter figures out how to generate the result

Imperative Programming
is like...

"Add 2 teaspoons of salt
and 2 teaspoons of
yeast.
Add 3 cups of flour.
Add 2 tablespoons of
olive oil.
Add 1/4th a cup of
water;
Start mixing the
ingredients together.
Put the dough ball on a
surface..."



Declarative Programming
is like...
"Make me a pizza"

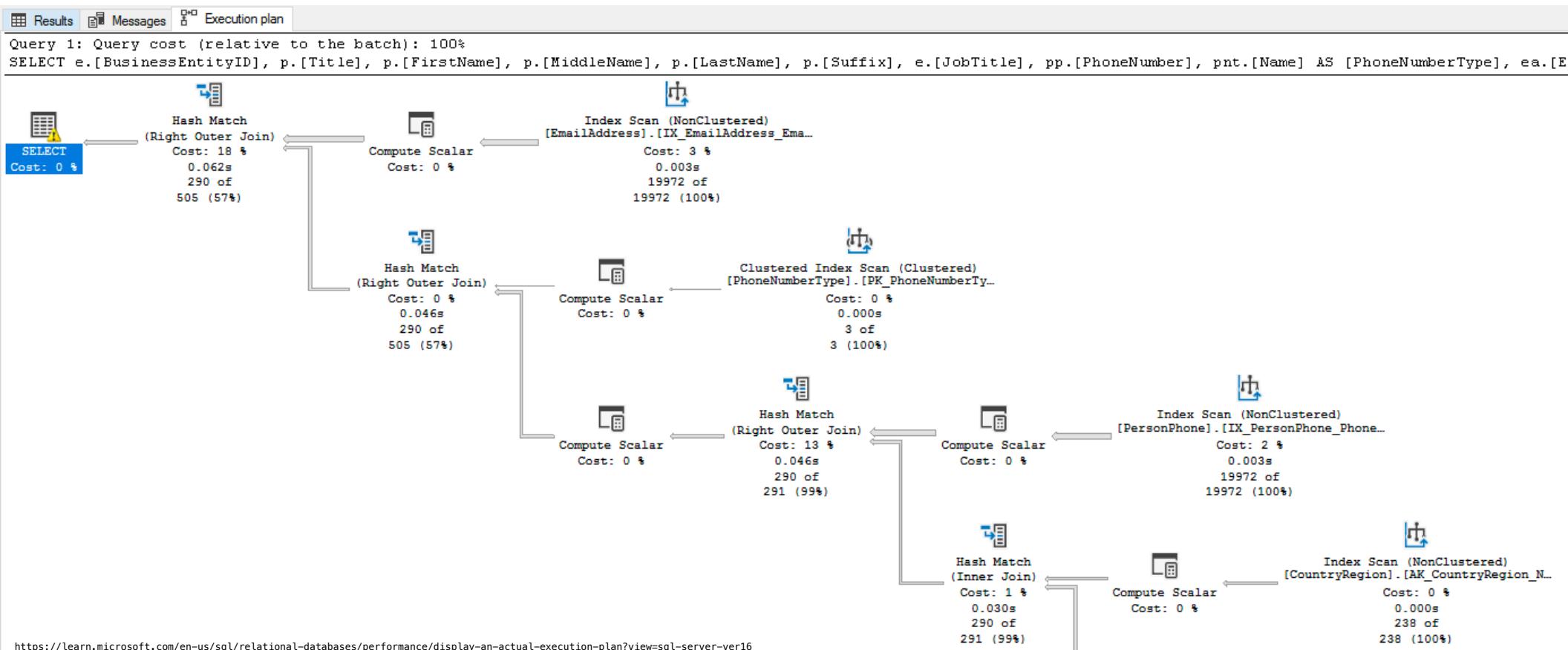


Declarative Programming

In **declarative programming**:

- A "program" is a description of the desired result
- The interpreter figures out how to generate the result

SQL Server Query Plan:



Structured Query Language (SQL)

Naming Tables

A **select** statement creates a new table and displays it.

A **create table** statement names the result of a **select** statement.

```
create table [name] as [select statement];
```

Here's how I might create a table of some of my most-listened-to spotify tracks in SQL:

```
create table spotify as  
  
select "WILDFLOWER" as track, "Billie Eilish" as artist union  
select "BIRDS OF A FEATHER" , "Billie Eilish" union  
select "360" , "Charli xcx" union  
select "Pasilyo" , "Sunkissed Lola" union  
select "Cinderella" , "Remi Wolf" union  
select "Good Luck Babe!" , "Chappell Roan" union  
select "Meow" , "Anamanaguchi";
```

spotify:

track	artist
WILDFLOWER	Billie Eilish
BIRDS...	Billie Eilish
360	Charli xcx
Pasilyo	Sunkissed Lola
Cinderella	Remi Wolf
Good Luck Babe!	Chappell Roan
Meow	Anamanaguchi

Select Statements Project Existing Tables

A **select** statement can specify an input table using a **from** clause

A subset of the rows of the input table can be selected using a **where** clause

An ordering over the remaining rows can be declared using an **order by** clause

Column descriptions determine how each input row is projected to a result row

```
select [expression] as [name], [expression] as [name], ...;  
select [columns] from [table] where [condition] order by [order];  
select track from spotify where artist = "Billie Eilish";  
select track from spotify where track < artist;
```

spotify:

track	artist
360	Charli xcx
BIRDS...	Billie Eilish
Cinderella	Remi Wolf
Good Luck Babe!	Chappell Roan
Pasilyo	Sunkissed Lola
Meow	Anamanaguchi
WILDFLOWER	Billie Eilish

Example: UC Salary Data, Passwords

SOURCES: <https://transparentcalifornia.com/download/salaries/university-of-california/>
[https://www.kaggle.com/datasets/babaralijamali/leakedpasswords?
select=top_200_password_2020_by_country.csv](https://www.kaggle.com/datasets/babaralijamali/leakedpasswords?select=top_200_password_2020_by_country.csv)

The University is a public institution, so it is supported to an extent by California taxpayers through an allocation by the state government. In the past, generous state support allowed UC Berkeley to operate while keeping costs to students low. While still an important revenue source, the state's financial support of the university has diminished significantly. Thirty years ago, 50 percent of the university's revenue came from the state, but today, the state provides just 14 percent of the university's revenue.

