Making an RDBMS Data Scientist Friendly

Advanced In-database Interactive Analytics with Visualization Support

McGill

Joseph Vinish D'Silva **Florestan De Moor** **Bettina Kemme**

{joseph.dsilva, florestan.demoor}@mail.mcgill.ca kemme@cs.mcgill.ca

VLDB 2019

Current in-database analytics approaches

Linear Algebra Using SQL

```
SELECT m1.i, m2.j, SUM(m1.v * m2.v)
FROM matrix AS m1, matrix AS m2
WHERE m1.j = m2.i
GROUP BY m1.i, m2.j;
```

- SQL is not intuitive for linear algebra.

HLL UDFs

```
CREATE FUNCTION lnrReg(…) LANGUAGE PYTHON
{
  //Read data from database tables.
  //HLL statements ,linear algebra, etc.
  //Save objects needed later into the db.
};
```

- Procedural syntax not suitable for exploratory work.

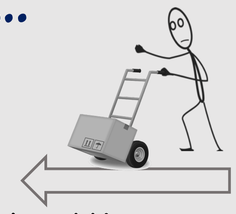
Performance

Usability

What data scientists do …

User System:

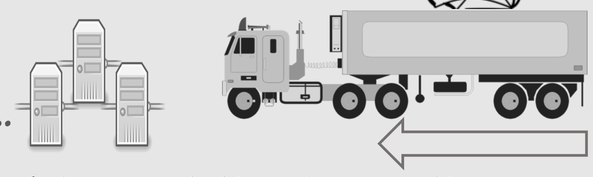
R, Python, pandas, …



- Less computing resources, data subsetting.

Big data Clusters:

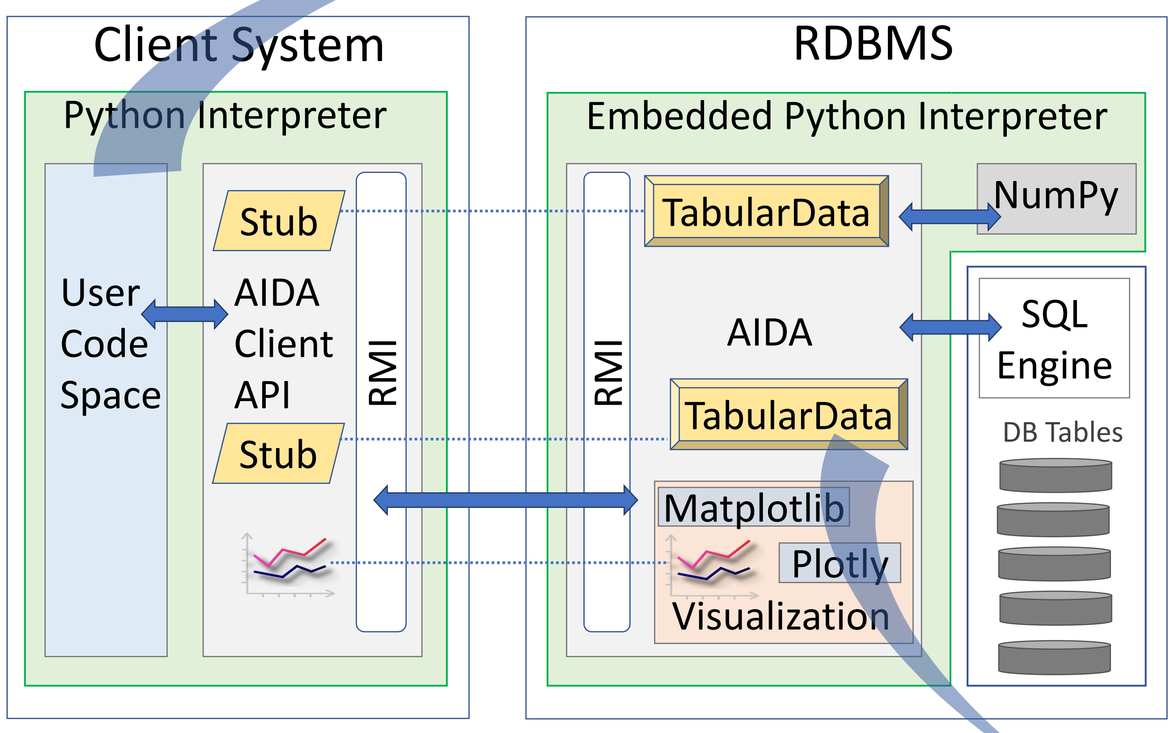
Hadoop, Spark, …



- Data transfer delays, relational operations are not efficient in these systems.

AIDA → Goals : Performance & Usability

Architecture

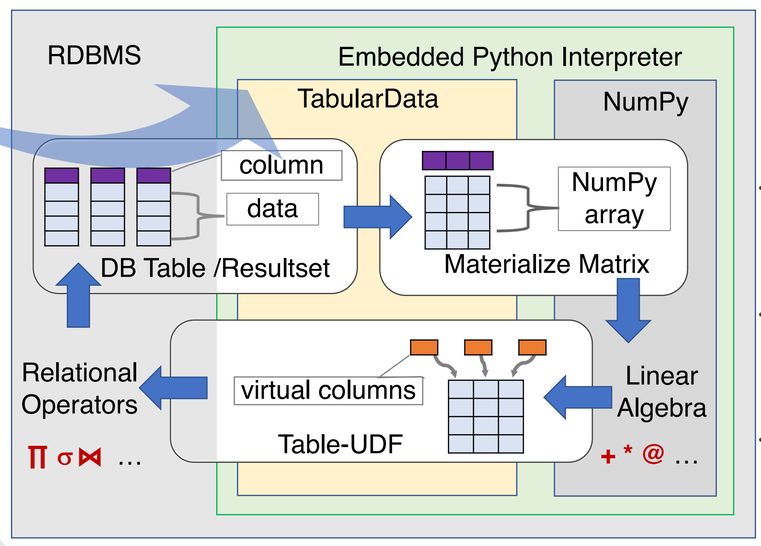


- All computation is performed inside the RDBMS.
- Client-server model, using RMI.
- TabularData objects support both relational and linear algebra operations.
- Relational operations follow ORM syntax, translated by AIDA to SQL, executed by RDBMS.
- Linear algebra executed using NumPy.
- Visualizations supported using Matplotlib and Plotly.

Usage

Client tool: Python interpreter / Jupyter notebook

```
//Establish a connection to AIDA server.
db = aida.connect(user='tpch', pass='...', ..)
ct = db.customer //Ref. to customer table in db.
//Find the number of customers in mkt. segments.
t1 = ct.agg(('c_mktsegment'
       ,{COUNT('*'): 'ncusts'}), ('c_mktsegment'))
//Total num. of customers, via matrix mul.
t2 = t1[['ncusts']].T @ t1[['ncusts']]
//See the actual results (ship to client).
print(t2.cdata) //Data size is small.
  ...
//mktseg_barchart is a user defined function
// that returns a Matplotlib or Plotly object.
db._Plot(mktseg_barchart, t1).
```

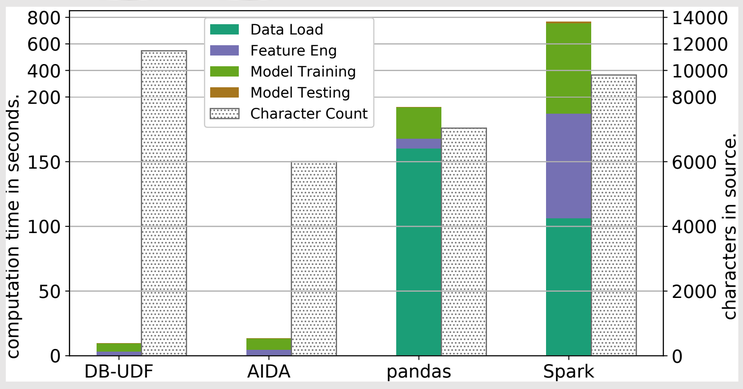


TabularData Unified Abstraction

- Supports interleaved linear algebra and relational operations.
- Zero-copy optimization minimizes data movement overheads.
- Table-UDFs are used to perform SQL on NumPy data structures.

System Comparisons

linear regression



Programming paradigms

	Languages	Interactive	Incremental	Near-data	Visualization	Unified
AIDA	Python	✓	✓	✓	✓	✓
DB UDF	Python, SQL	✗	✗	✓	✗	✓
pandas	Python	✓	✓	✗	✓	✗
Spark	Scala	✓	✓	✗	✓	✗

[1] JV D'Silva, F De Moor, B Kemme. *Making an RDBMS Data Scientist Friendly: Advanced In-database Interactive Analytics with Visualization Support.* PVLDB, 12(12):1930-1933 , 2019.
[2] JV D'Silva, F De Moor, B Kemme. *AIDA – Abstraction for Advanced In-database Analytics.* PVLDB, 11(11):1400-1413 , 2019.
[3] JV D'Silva, F De Moor, B Kemme. *Keep Your Host Language Object and Also Query it: A Case for SQL Query Support in RDBMS for Host Language Objects.* SSDBM, 2019.