AI Accelerator: Overcome Data Quality/Wrangling

Overview [Distributed SQL + Full Fidelity Metadata]

Demonstration

(September 2020)

Data Challenges

Data Challenges

Poor data quality costs an average customer $15M per year

Gartner (2018)

"How to Stop Data Quality Undermining Your Business"

Data Challenges

Poor data quality costs an average customer $15M per year

Gartner (2018)

"How to Stop Data Quality Undermining Your Business"

2/3rd of developers don't believe they can deliver trusted data

IDC (2019)

"Most Enterprises Don't Trust Their Data, According to New IDC/Talend Survey"

Data Challenges

Poor data quality costs an average customer $15M per year

Gartner (2018)

"How to Stop Data Quality Undermining Your Business"

2/3rd of developers don't believe they can deliver trusted data

IDC (2019)

"Most Enterprises Don't Trust Their Data, According to New IDC/Talend Survey"

80% time of AI projects is spent on collecting and preparing data

Wall Street Journal (2019)

"Data Challenges Are Halting AI Projects"

Data Challenges

Poor data quality costs an average customer $15M per year

Gartner (2018)

"How to Stop Data Quality Undermining Your Business"

2/3rd of developers don't believe they can deliver trusted data

IDC (2019)

"Most Enterprises Don't Trust Their Data, According to New IDC/Talend Survey"

80% time of AI projects is spent on collecting and preparing data

Wall Street Journal (2019)

"Data Challenges Are Halting AI Projects"

Data challenges consume large portion of customer AI/Analytics budgets

Profile Metadata is Critical for AI/Analytics

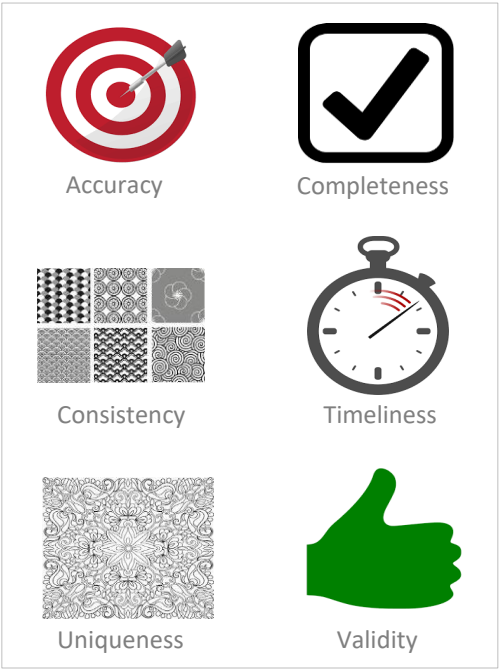
Profile Metadata is Critical for AI/Analytics

Profile metadata = Statistics + informational summaries about dataset hygiene & quality

Profile Metadata is Critical for AI/Analytics

Profile metadata = Statistics + informational summaries about dataset hygiene & quality

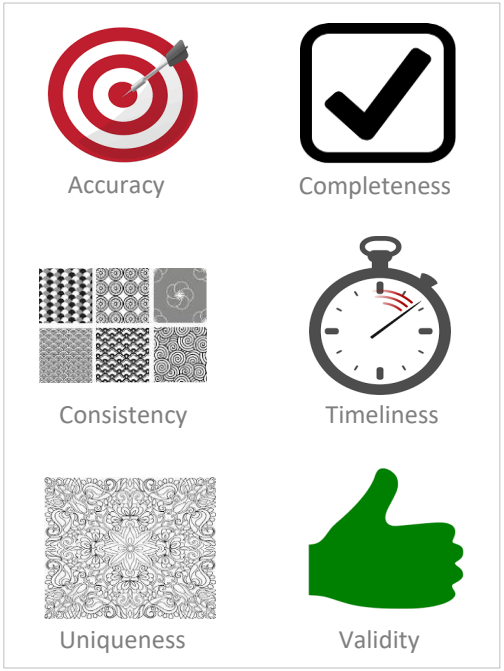
Metadata insights



Profile Metadata is Critical for AI/Analytics

Profile metadata = Statistics + informational summaries about dataset hygiene & quality

Metadata insights



Profiling is tedious

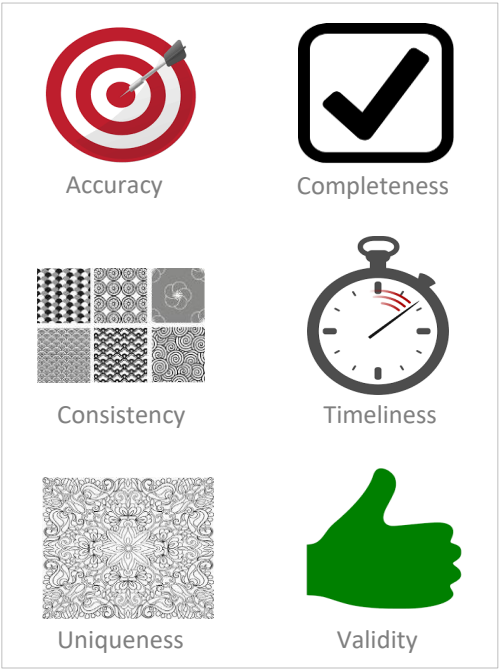
"coffee, lunch, overnight"

Felix Naumann
Hasso Platner Institute
"An Introduction to Data Profiling"

Profile Metadata is Critical for AI/Analytics

Profile metadata = Statistics + informational summaries about dataset hygiene & quality

Metadata insights



Profiling is tedious

"coffee, lunch, overnight"

Felix Naumann
Hasso Platner Institute
"An Introduction to Data Profiling"

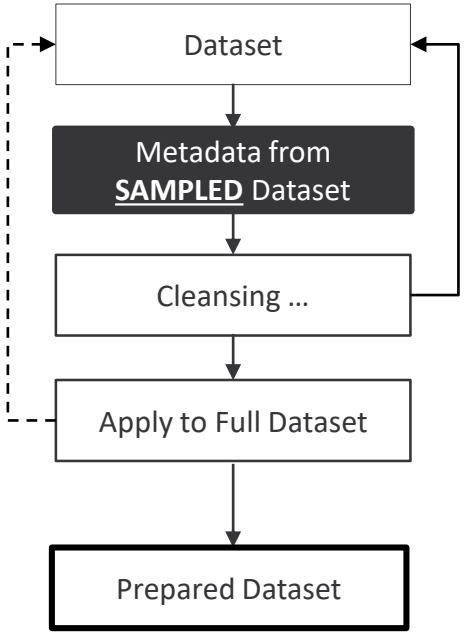
Industrywide practice

Generate metadata from sampled data

A Paradigm Shift to Overcome Data Challenges

A Paradigm Shift to Overcome Data Challenges

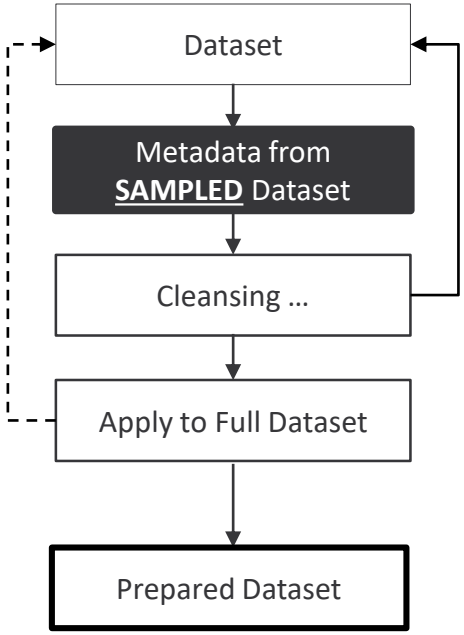
Industry Practice



- ✗ Piecemeal metadata insights
- ✗ Iterative process
- ✗ Labor & time intensive

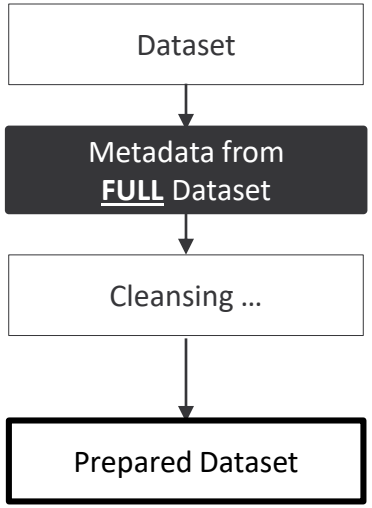
A Paradigm Shift to Overcome Data Challenges

Industry Practice



- ✗ Piecemeal metadata insights
- ✗ Iterative process
- ✗ Labor & time intensive

XtremeData's Transformative Solution



- ✓ Complete metadata insights
- ✓ Eliminates iterations
- ✓ 60-80% productivity increase

XtremeData: Engine to Generate Full Fidelity Metadata

XtremeData: Engine to Generate Full Fidelity Metadata

Unparalleled Simplicity

Full dataset profiling: SQL-enabled, fast, scalable …

Column level profiling, including schema inferencing

Rich metadata: data quality, patterns, ranks, MCVs …

Uniform structure for profile results: easily exported, or queried for discovery, curation …

XtremeData: Engine to Generate Full Fidelity Metadata

Unparalleled Simplicity

Full dataset profiling: SQL-enabled, fast, scalable …

Column level profiling, including schema inferencing

Rich metadata: data quality, patterns, ranks, MCVs …

Uniform structure for profile results: easily exported, or queried for discovery, curation …

Unparalleled Price-Performance

Records[1]	Dataset	Time (sec)[2]	Cost[3]
2B	0.53TB	331	$2.43
4B	1.06TB	635	$4.67
8B	2.1TB	1285	$9.45
16B	4.2TB	2609	$19.18

[1] 81 columns (all varchar 64) [2] System: 6x c5.9xlarge (AWS) [3] Cost: AWS Compute + XtremeData

XtremeData: Engine to Generate Full Fidelity Metadata

Unparalleled Simplicity

Full dataset profiling: SQL-enabled, fast, scalable …

Column level profiling, including schema inferencing

Rich metadata: data quality, patterns, ranks, MCVs …

Uniform structure for profile results: easily exported, or queried for discovery, curation …

Unparalleled Price-Performance

Records[1]	Dataset	Time (sec)[2]	Cost[3]
2B	0.53TB	331	$2.43
4B	1.06TB	635	$4.67
8B	2.1TB	1285	$9.45
16B	4.2TB	2609	$19.18

[1] 81 columns (all varchar 64) [2] System: 6x c5.9xlarge (AWS) [3] Cost: AWS Compute + XtremeData

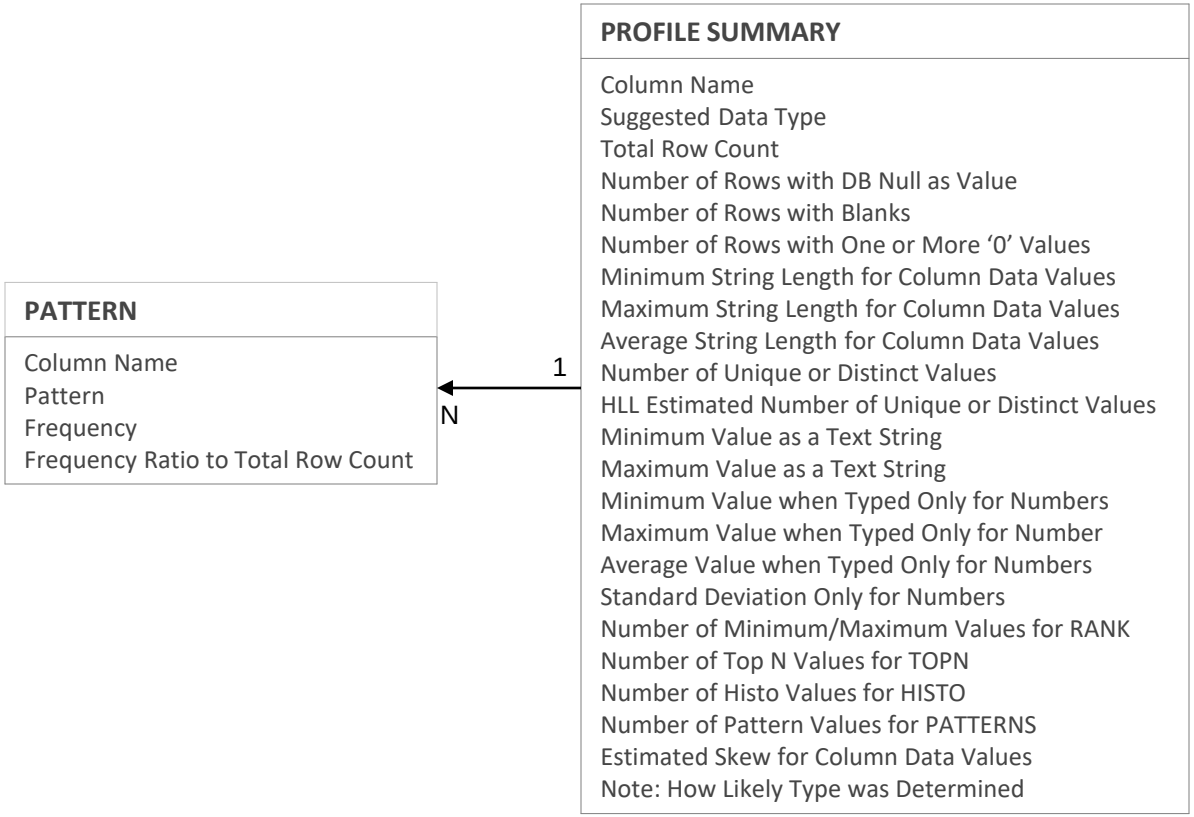
Enables enterprises to generate complete metadata from data at-rest & in-motion

Profile Metadata Organized for Efficient Access

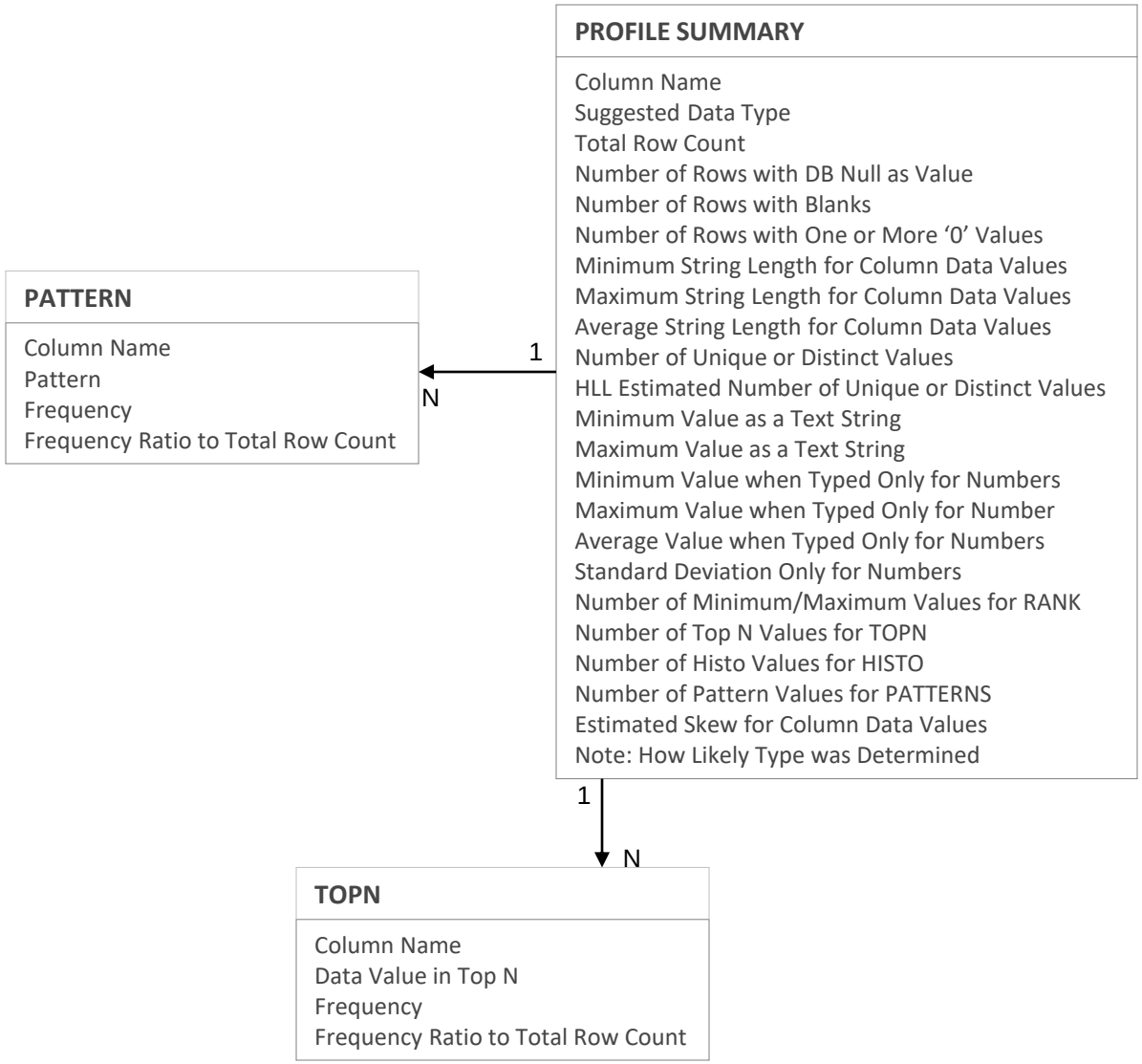
Profile Metadata Organized for Efficient Access

PROFILE SUMMARY
Column Name
Suggested Data Type
Total Row Count
Number of Rows with DB Null as Value
Number of Rows with Blanks
Number of Rows with One or More ‘0’ Values
Minimum String Length for Column Data Values
Maximum String Length for Column Data Values
Average String Length for Column Data Values
Number of Unique or Distinct Values
HLL Estimated Number of Unique or Distinct Values
Minimum Value as a Text String
Maximum Value as a Text String
Minimum Value when Typed Only for Numbers
Maximum Value when Typed Only for Number
Average Value when Typed Only for Numbers
Standard Deviation Only for Numbers
Number of Minimum/Maximum Values for RANK
Number of Top N Values for TOPN
Number of Histo Values for HISTO
Number of Pattern Values for PATTERNS
Estimated Skew for Column Data Values
Note: How Likely Type was Determined

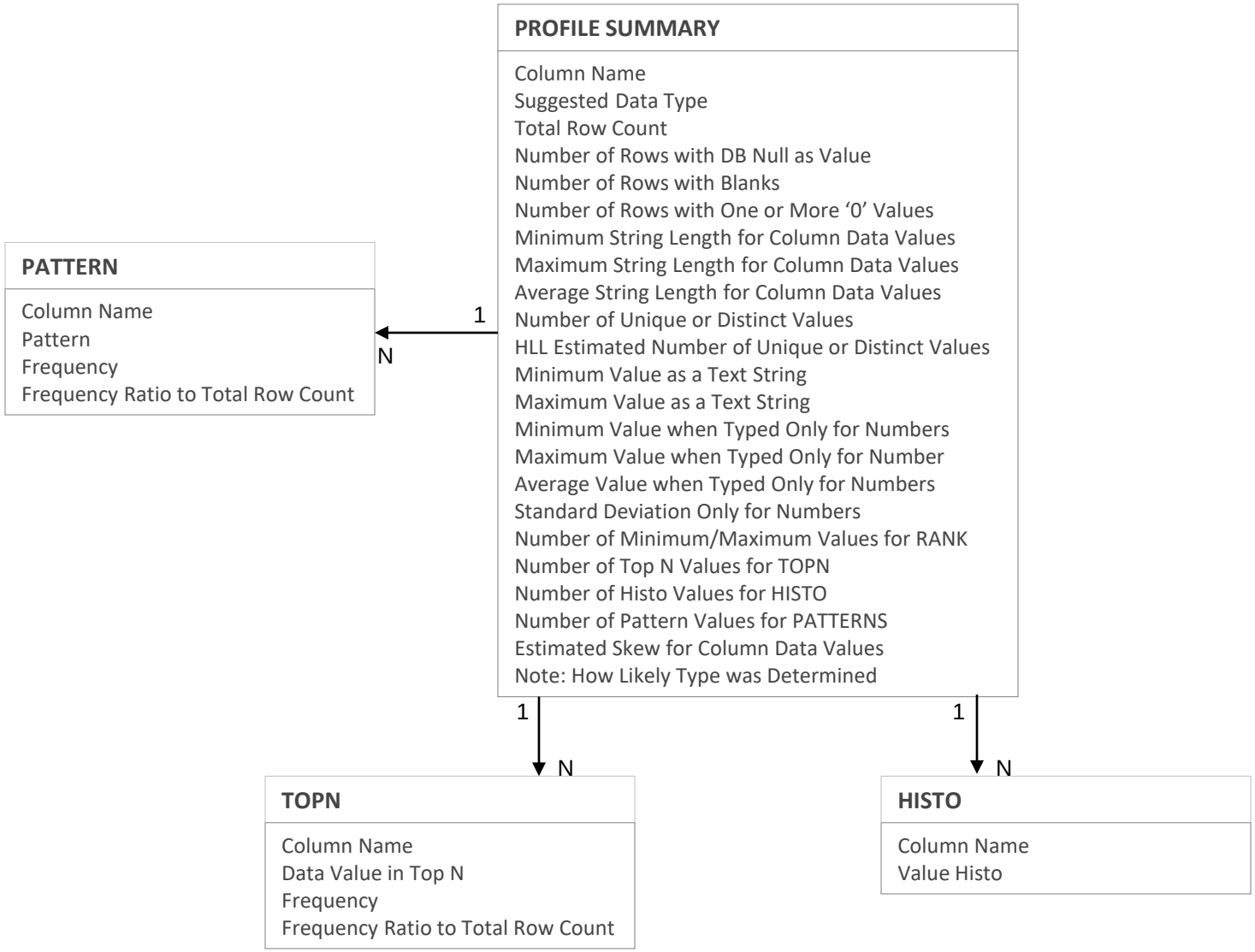
Profile Metadata Organized for Efficient Access



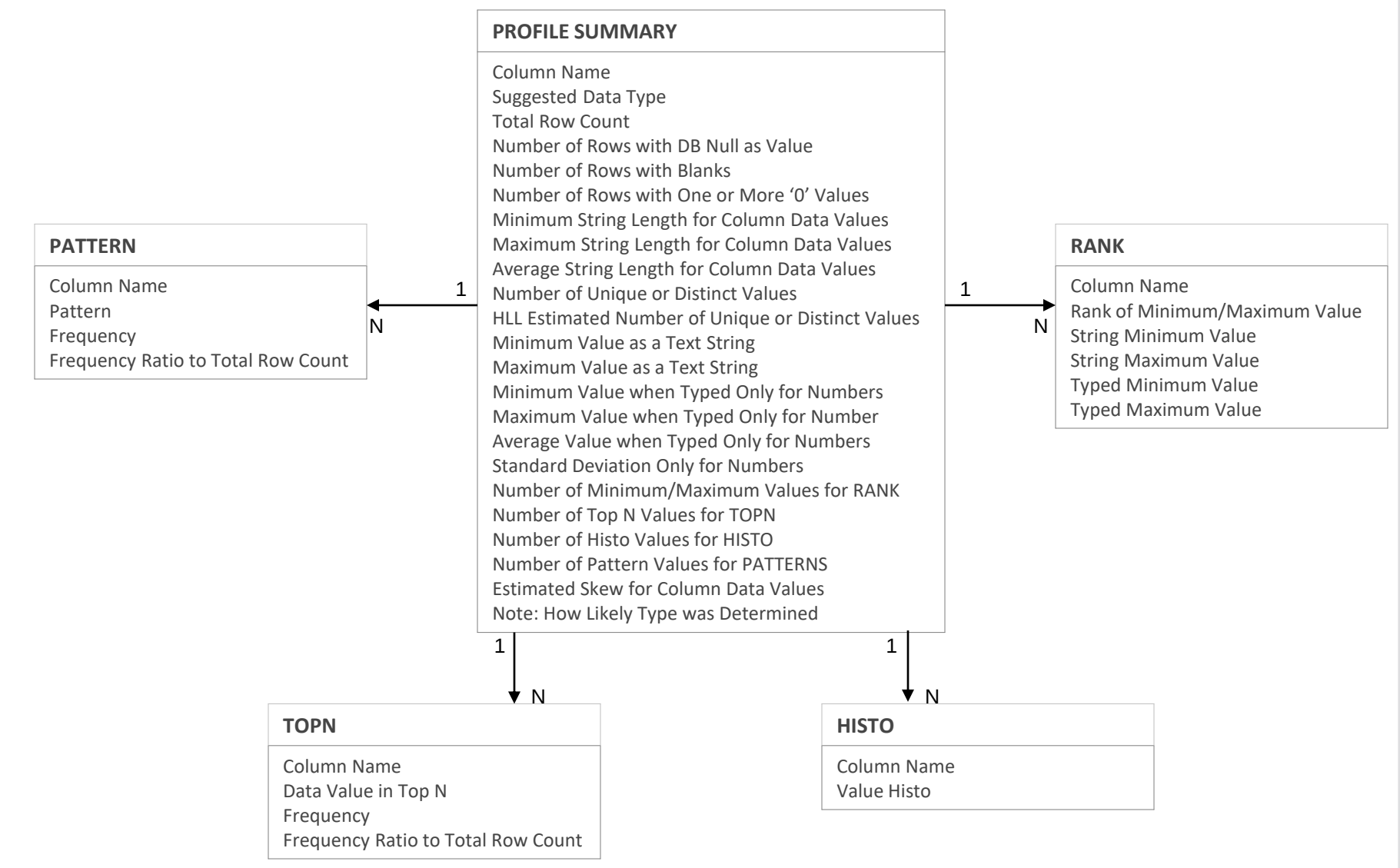
Profile Metadata Organized for Efficient Access



Profile Metadata Organized for Efficient Access



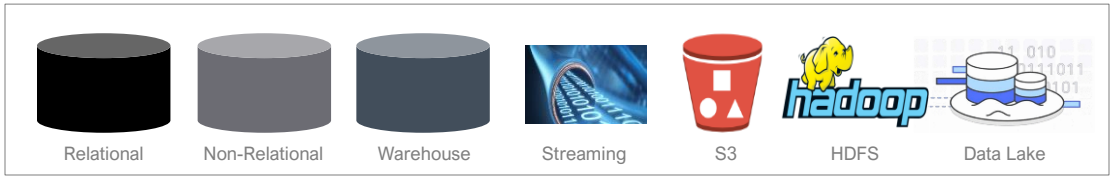
Profile Metadata Organized for Efficient Access



Benefits of Full Dataset Profiling (FDP)

Benefits of Full Dataset Profiling (FDP)

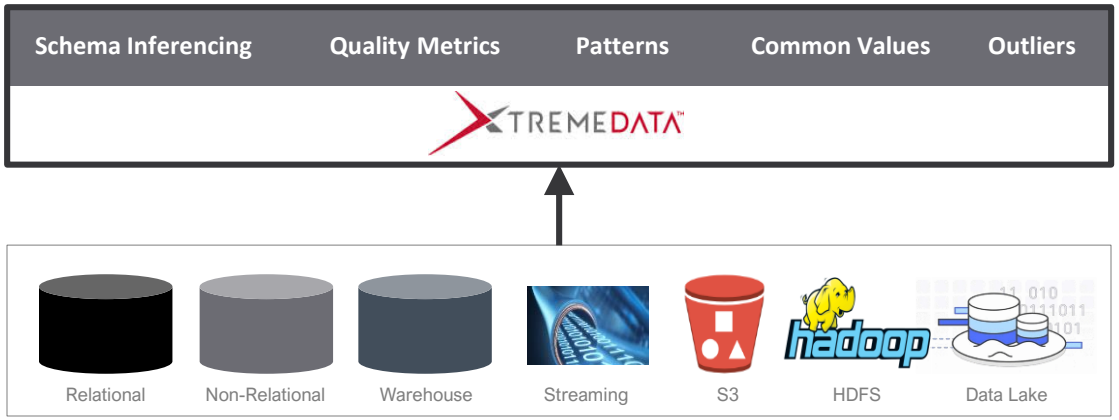
TB-PBs of data at-rest & in-motion



Benefits of Full Dataset Profiling (FDP)

Build complete metadata for all data

TB-PBs of data at-rest & in-motion

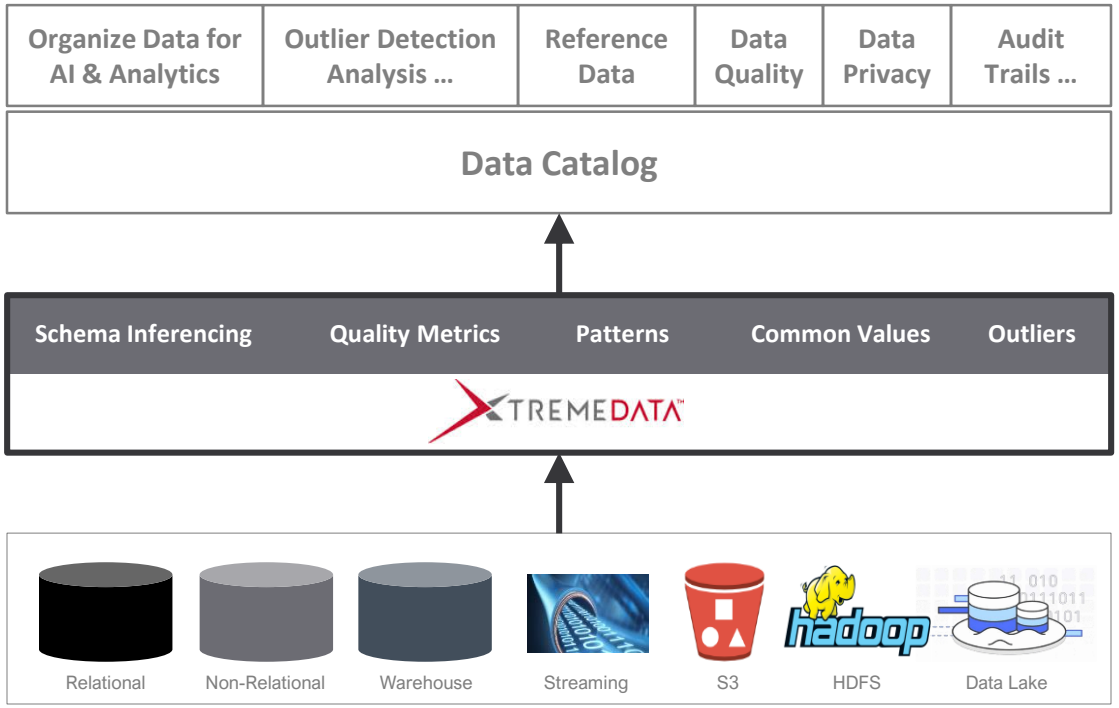


Benefits of Full Dataset Profiling (FDP)

Metadata can be stored in data catalogs

Build complete metadata for all data

TB-PBs of data at-rest & in-motion

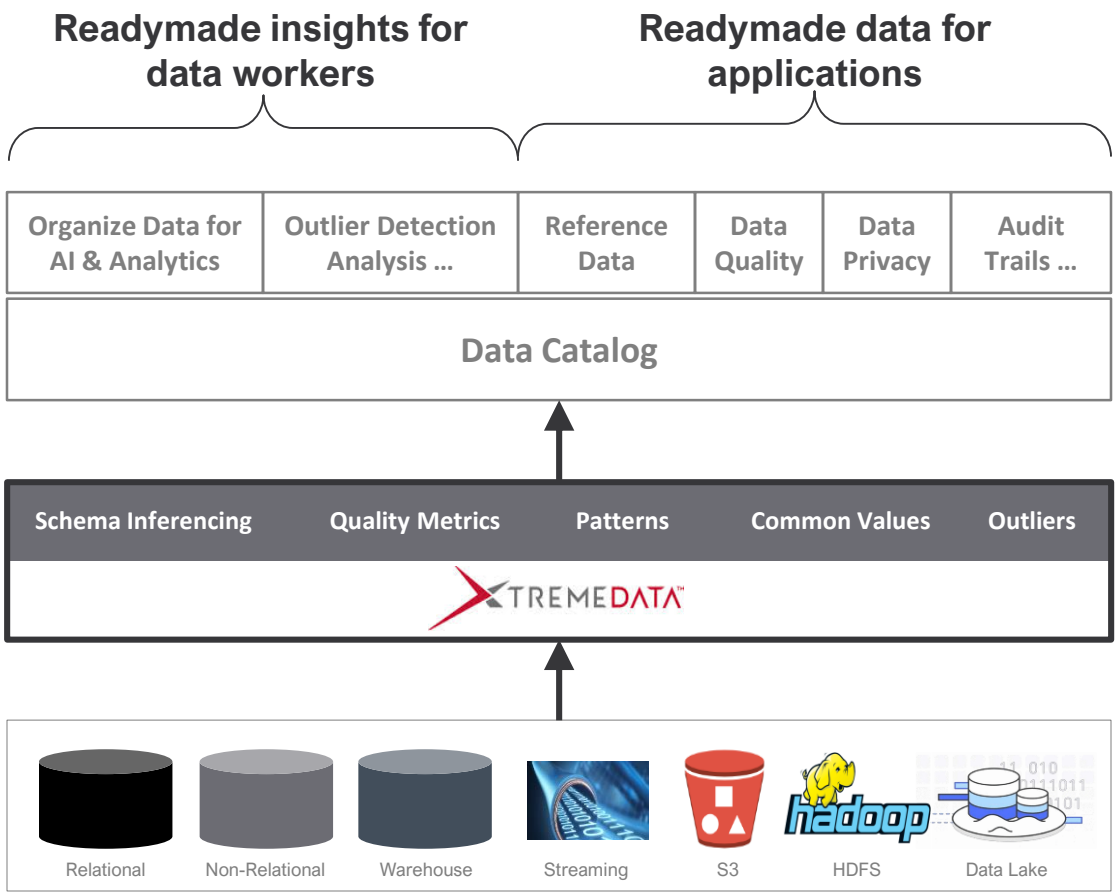


Benefits of Full Dataset Profiling (FDP)

Metadata can be stored in data catalogs

Build complete metadata for all data

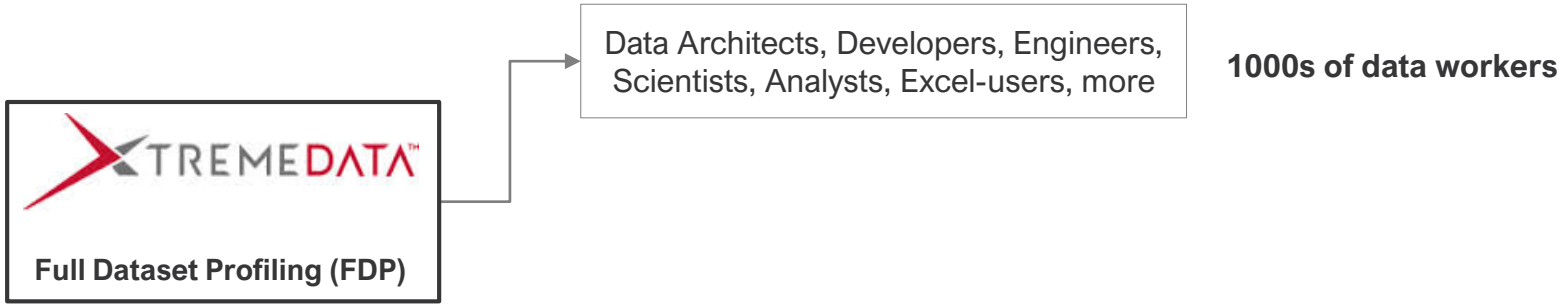
TB-PBs of data at-rest & in-motion



Compelling Customer ROI

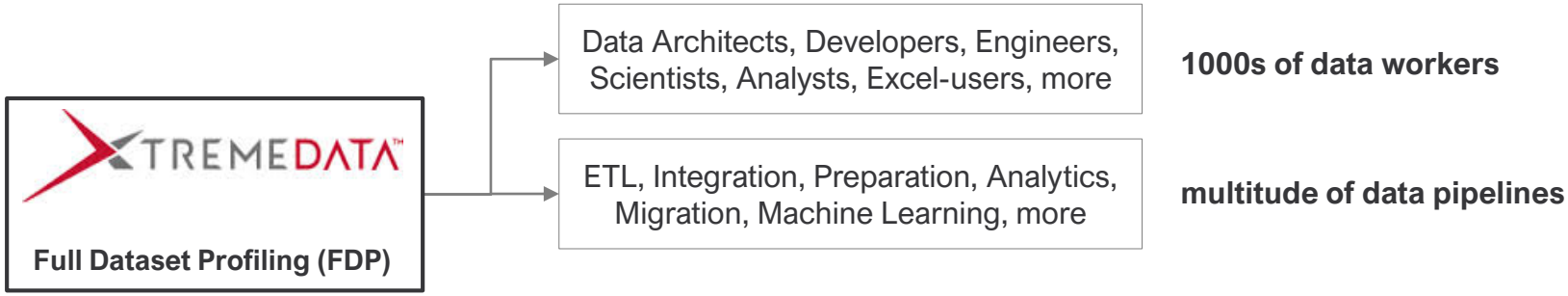
Compelling Customer ROI

Use in a typical enterprise



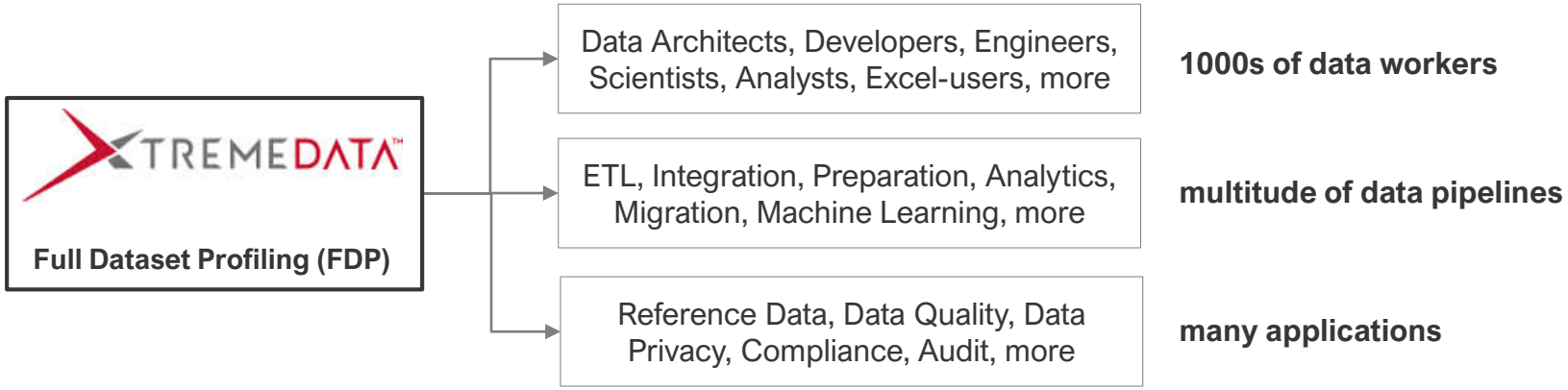
Compelling Customer ROI

Use in a typical enterprise



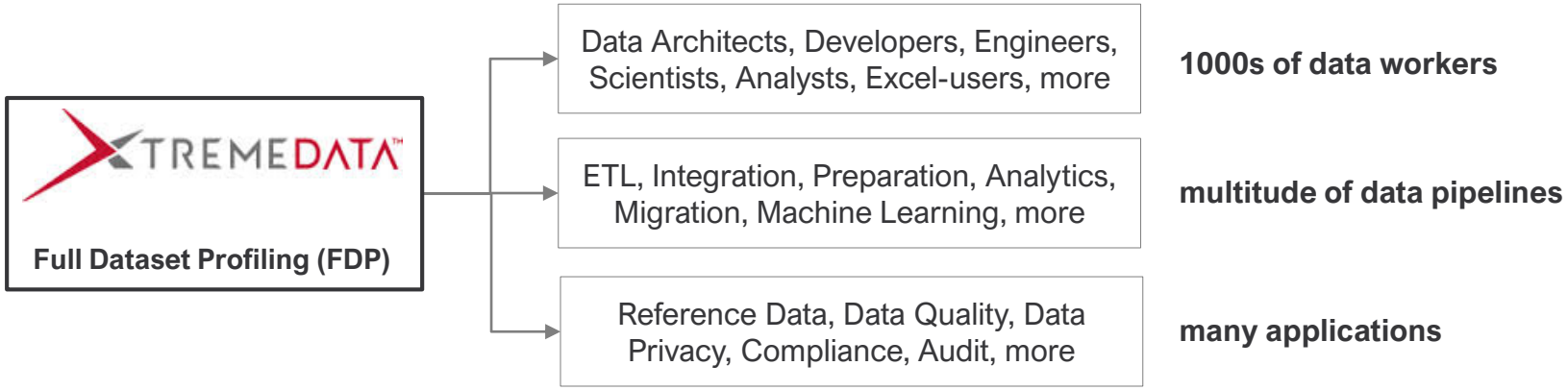
Compelling Customer ROI

Use in a typical enterprise



Compelling Customer ROI

Use in a typical enterprise

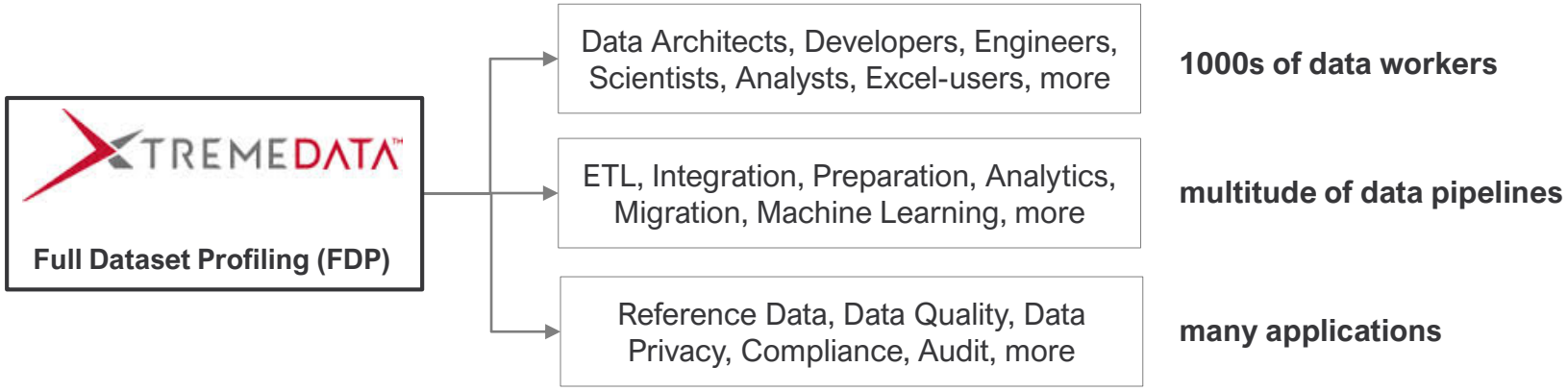


Few dollars in FDP provide *data & insights* that can save tens of thousands

30% reduction in data wrangling for a small team of 4 @ $75/hour/employee saves $180K per year

Compelling Customer ROI

Use in a typical enterprise



Few dollars in FDP provide *data & insights* that can save tens of thousands

30% reduction in data wrangling for a small team of 4 @ $75/hour/employee saves <u>$180K per year</u>

Customer with 1000 data workers @ $100m ($100k/worker) & FDP[1] @ $105k

Modest 15% productivity gain: $15M

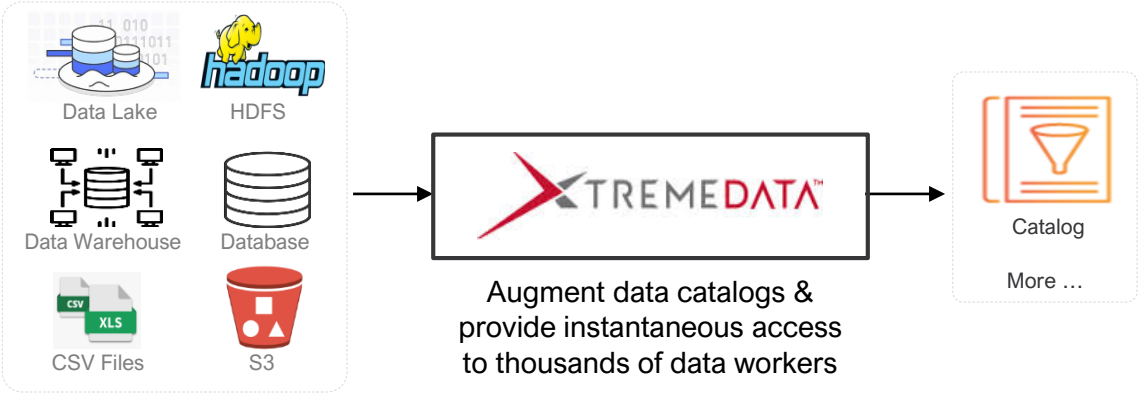
Potential ROI: <u>>140x</u>

[1] System: 8x c3.8xlarge (AWS) + XtremeData, running 12 hours x 5 days x 52 weeks

Use Cases

Use Cases

Faster cloud migrations and data science



Use Cases

Faster cloud migrations and data science

Augment data catalogs & provide instantaneous access to thousands of data workers

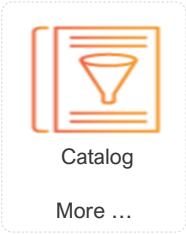
Complete metadata insights

- Data quality assessment
- Precise data structures (type, sale, precision)
- Facilitates automation
- Faster data science (data prep)
- Faster migration to cloud

Use Cases

Faster cloud migrations and data science

Augment data catalogs & provide instantaneous access to thousands of data workers

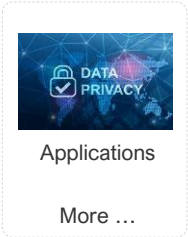


Complete metadata insights

- Data quality assessment
- Precise data structures (type, sale, precision)
- Facilitates automation
- Faster data science (data prep)
- Faster migration to cloud

Avoid fines for data privacy & protection

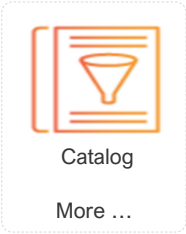
Enable continuous data privacy and protection



Use Cases

Faster cloud migrations and data science

Augment data catalogs & provide instantaneous access to thousands of data workers

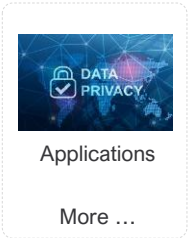


Complete metadata insights

- Data quality assessment
- Precise data structures (type, sale, precision)
- Facilitates automation
- Faster data science (data prep)
- Faster migration to cloud

Avoid fines for data privacy & protection

Enable continuous data privacy and protection



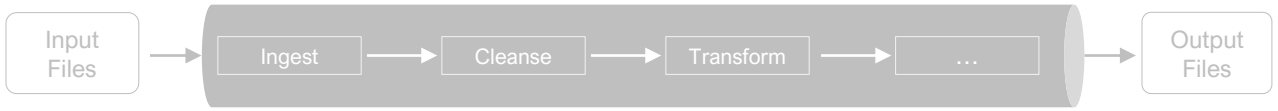
Metadata for data privacy

- All patterns (SSN, DoB, CC#)
- Stored in uniform structure
- Fast & easy discovery analytics
- Identify/protect data at any scale
- Readymade data for privacy & protection applications

Use Cases

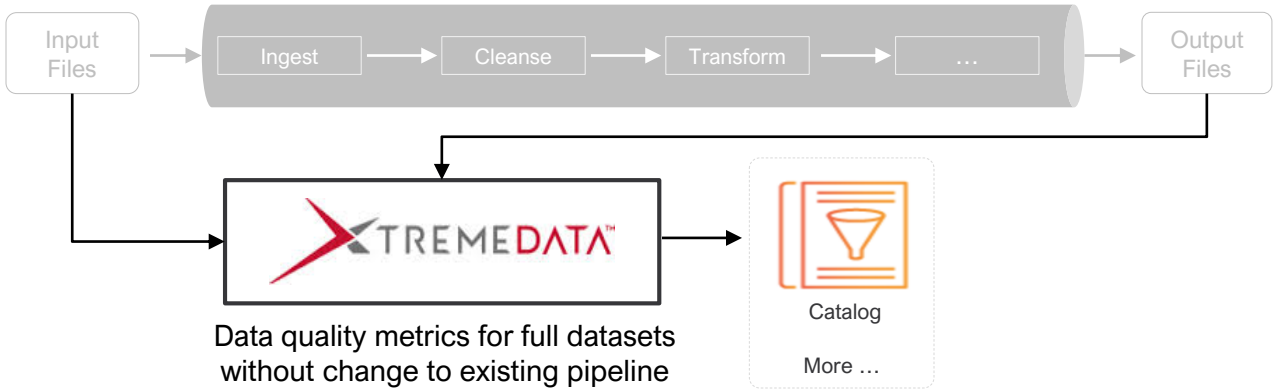
Use Cases

Reduce millions $s in operational costs & risks with data quality



Use Cases

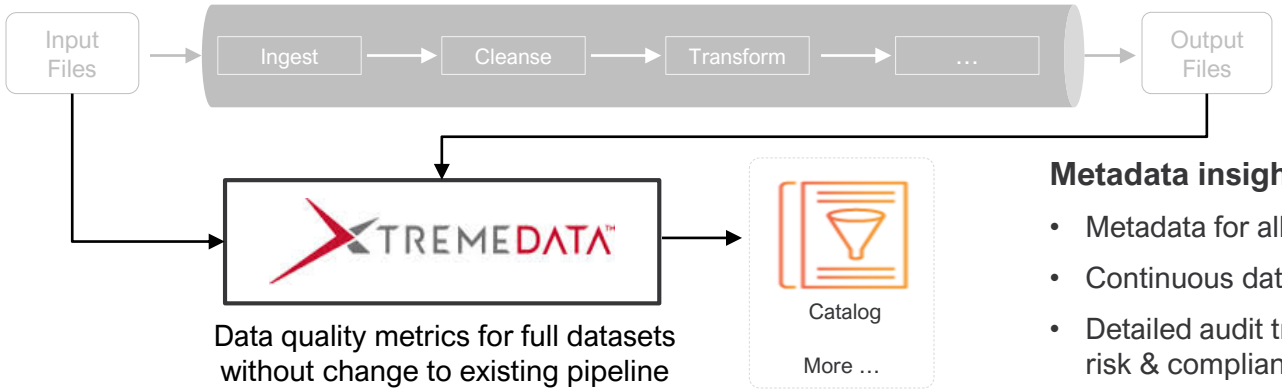
Reduce millions $s in operational costs & risks with data quality



Data quality metrics for full datasets without change to existing pipeline

Use Cases

Reduce millions $s in operational costs & risks with data quality



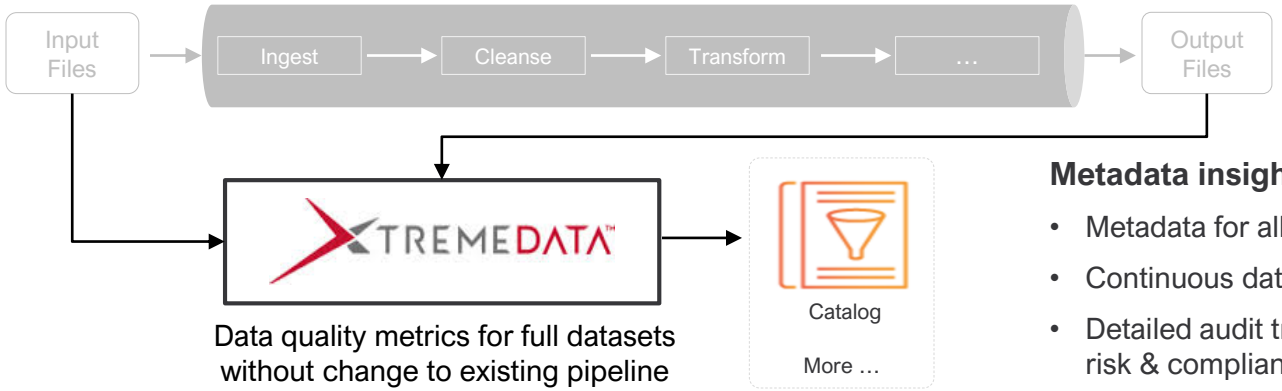
Data quality metrics for full datasets without change to existing pipeline

Metadata insights for data pipelines

- Metadata for all inputs/outputs
- Continuous data quality
- Detailed audit trail to enhance governance, risk & compliance (GRC) applications

Use Cases

Reduce millions $s in operational costs & risks with data quality

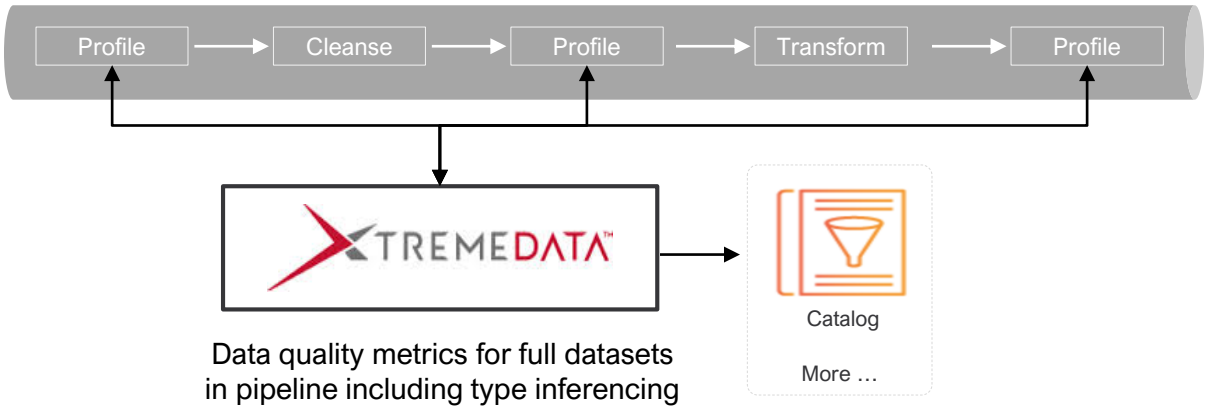


Data quality metrics for full datasets without change to existing pipeline

Metadata insights for data pipelines

- Metadata for all inputs/outputs
- Continuous data quality
- Detailed audit trail to enhance governance, risk & compliance (GRC) applications

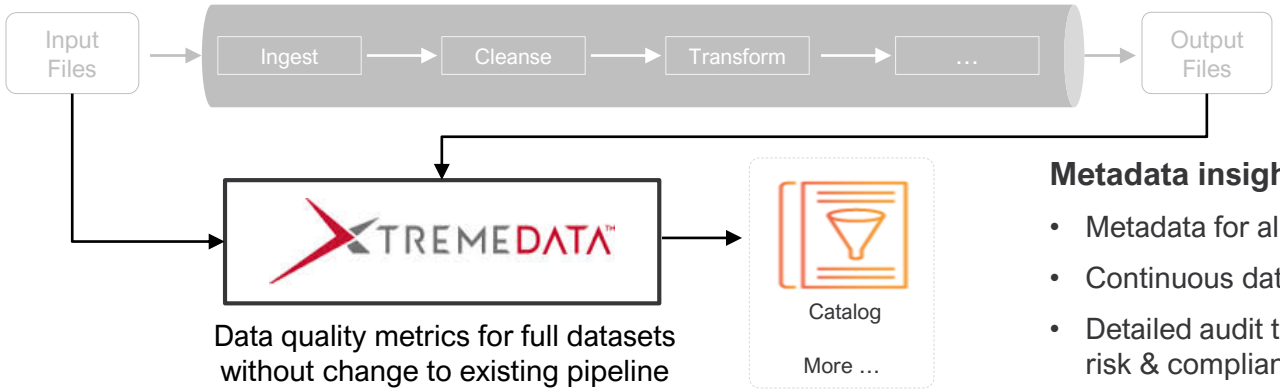
Build trust in data with transparency



Data quality metrics for full datasets in pipeline including type inferencing

Use Cases

Reduce millions $s in operational costs & risks with data quality

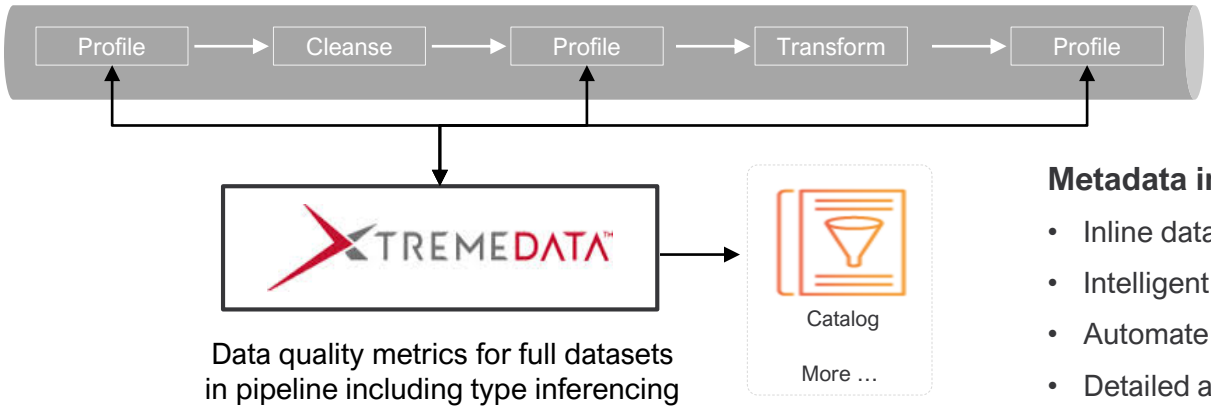


Data quality metrics for full datasets without change to existing pipeline

Metadata insights for data pipelines

- Metadata for all inputs/outputs
- Continuous data quality
- Detailed audit trail to enhance governance, risk & compliance (GRC) applications

Build trust in data with transparency



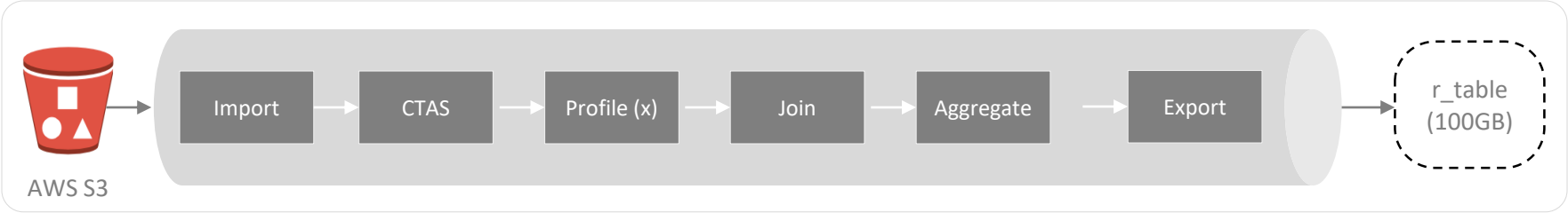
Data quality metrics for full datasets in pipeline including type inferencing

Metadata insights in data pipelines

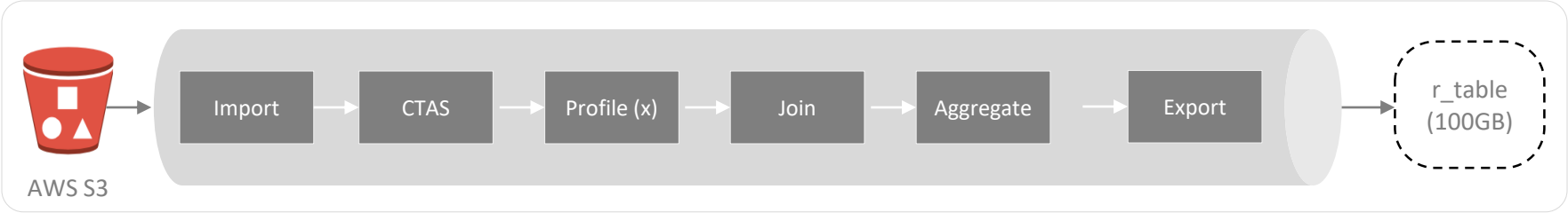
- Inline data quality + data transparency = trust
- Intelligent pipelines (detect change/drift …)
- Automate validation, reconciliation …
- Detailed audit trails for GRC applications

Demo

Demo


AWS S3
Import
CTAS
Profile (x)
Join
Aggregate
Export
r_table
(100GB)

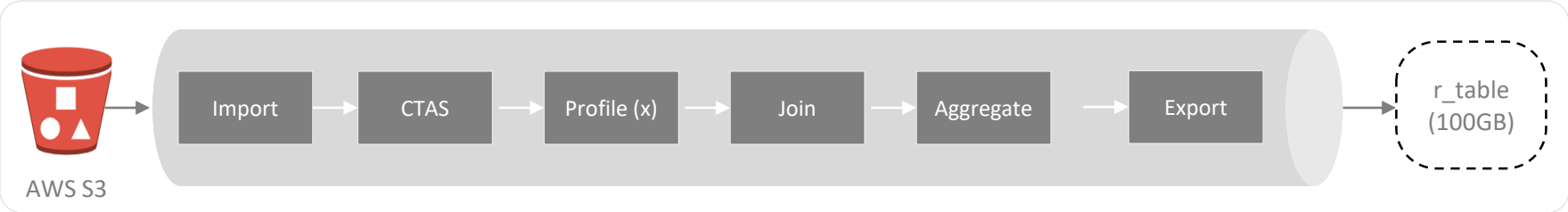
Demo



Market data (0.25TB)

- ASCII, CSV, Gzip
- ~50 files/table
- 5 tables
- 3.3 billion records

Demo



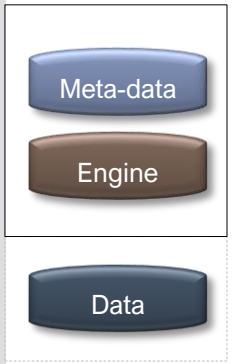
Market data (0.25TB)

- ASCII, CSV, Gzip
- ~50 files/table
- 5 tables
- 3.3 billion records

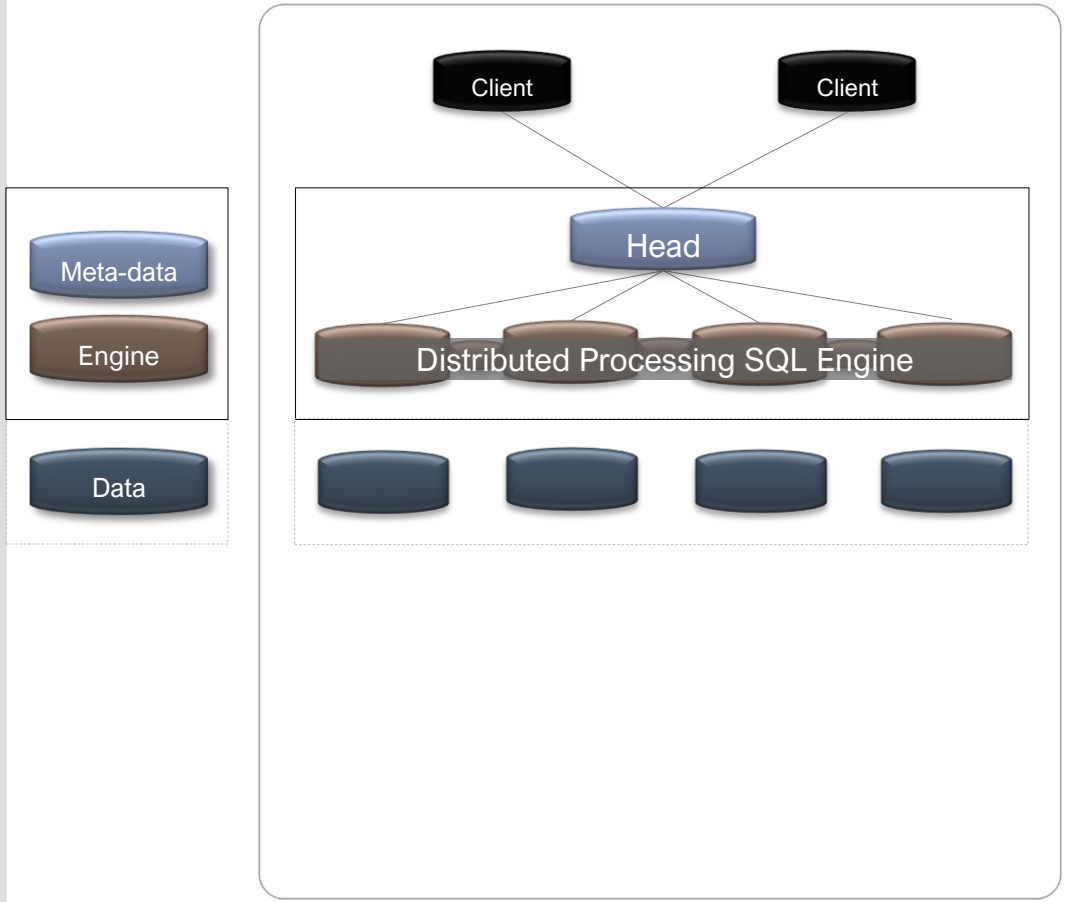
Import	c_table, o_table, q_table, r_table & t_table
Profile	Single column (symbol from r_table)
	All columns in table (c_table)
	All columns in table including data type inferencing (q_allchars)
Query	**Join**: t_table, r_table & o_table …
	Group: q_table …
Export	r_table

Architecture

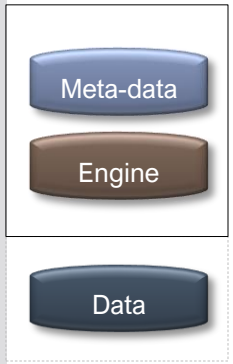
Architecture

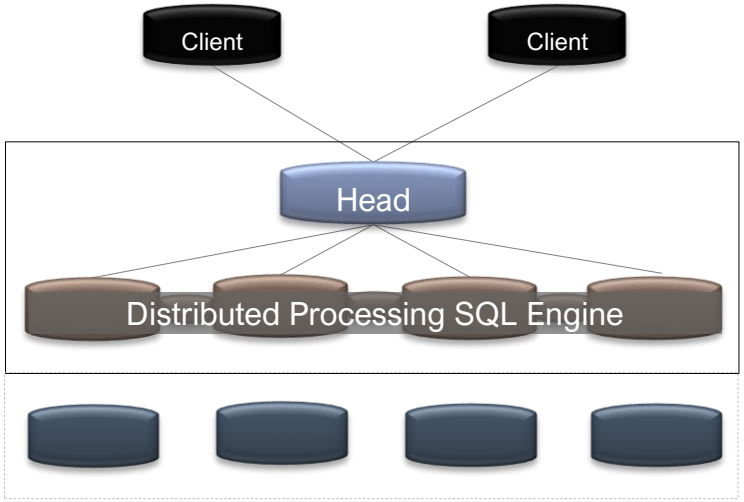


Architecture



Architecture

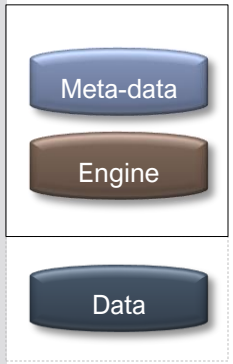


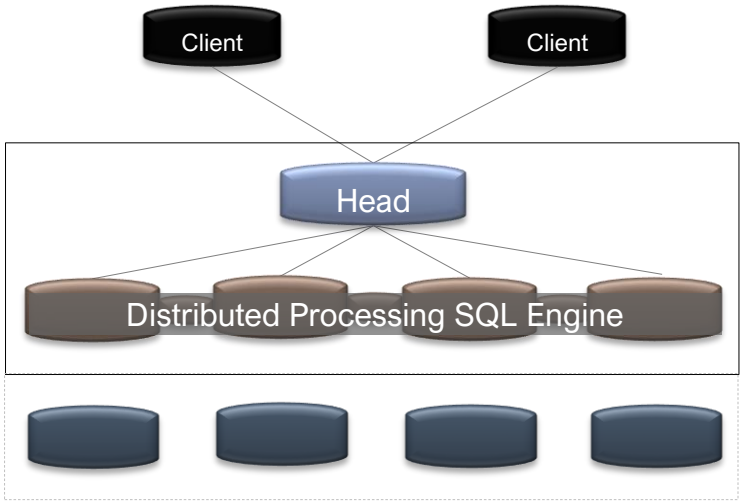


Single instance, peer-to-peer:

- Natively parallel: across & within data nodes
- Hybrid row & columnar processing
- Modern columnar, optimal auto-compression
- Micro optimized, multi-threaded, JIT ….

Architecture



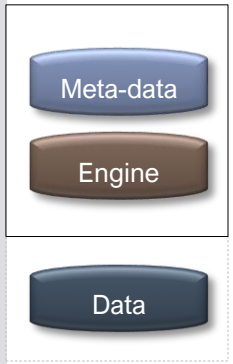


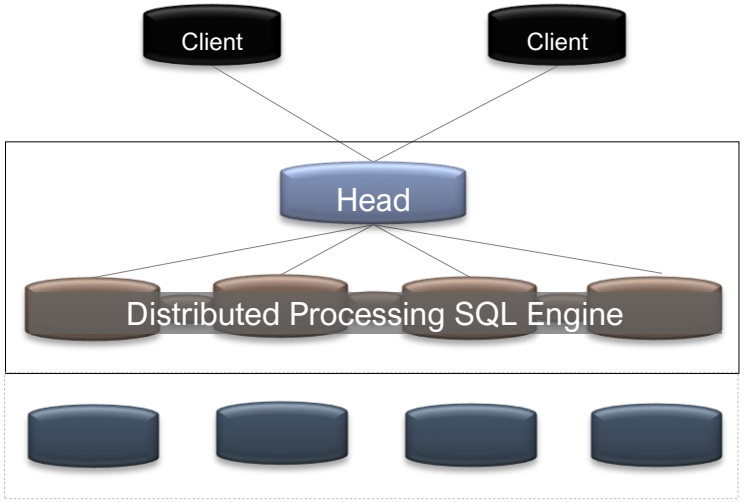
Single instance, peer-to-peer:

- Natively parallel: across & within data nodes
- Hybrid row & columnar processing
- Modern columnar, optimal auto-compression
- Micro optimized, multi-threaded, JIT ….

Wire protocol compatible with PostgreSQL
MPP extensions optional

Architecture





Single instance, peer-to-peer:

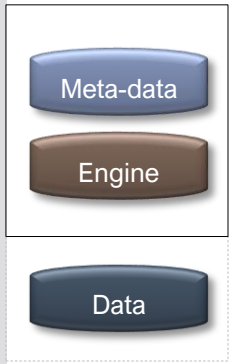
- Natively parallel: across & within data nodes
- Hybrid row & columnar processing
- Modern columnar, optimal auto-compression
- Micro optimized, multi-threaded, JIT ….

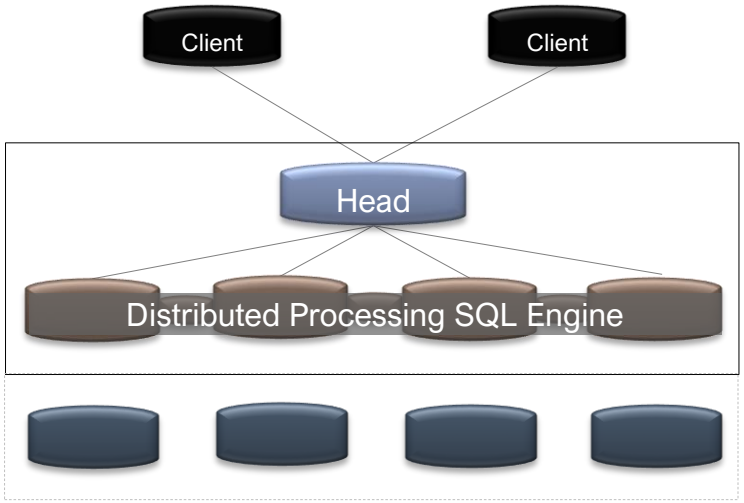
Wire protocol compatible with PostgreSQL
MPP extensions optional

Performance not dependent on data co-location

No secondary structures (indexes, zone maps ..)

Architecture





Single instance, peer-to-peer:

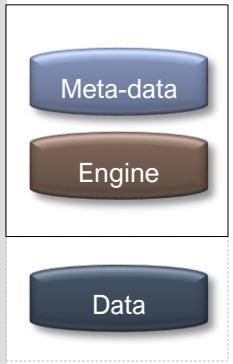
- Natively parallel: across & within data nodes
- Hybrid row & columnar processing
- Modern columnar, optimal auto-compression
- Micro optimized, multi-threaded, JIT ….

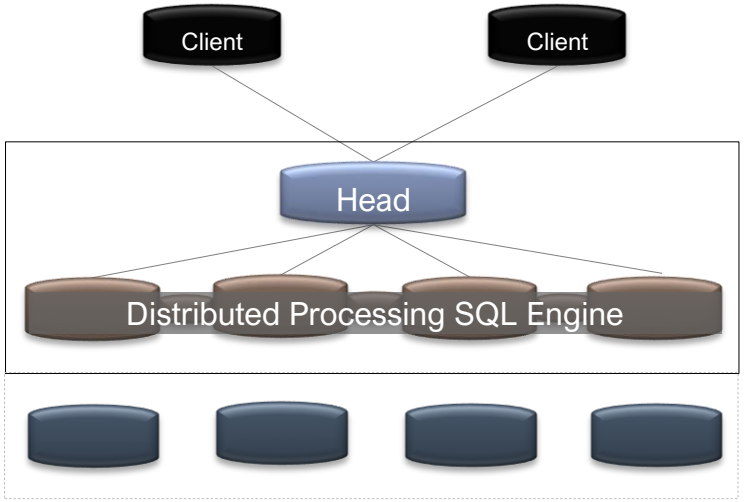
Wire protocol compatible with PostgreSQL
MPP extensions optional

Performance not dependent on data co-location
No secondary structures (indexes, zone maps ..)

Planner/optimizer with runtime plan refinement
Dynamic load balancing
Dynamic skew detection and handling

Architecture





Single instance, peer-to-peer:

- Natively parallel: across & within data nodes
- Hybrid row & columnar processing
- Modern columnar, optimal auto-compression
- Micro optimized, multi-threaded, JIT ….

Wire protocol compatible with PostgreSQL
MPP extensions optional

Performance not dependent on data co-location
No secondary structures (indexes, zone maps ..)

Planner/optimizer with runtime plan refinement
Dynamic load balancing
Dynamic skew detection and handling

Decoupled compute & data storage
Scale up & out

"Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the Networking and Information Technology Research and Development Program."

The Networking and Information Technology Research and Development (NITRD) Program

Mailing Address: NCO/NITRD, 2415 Eisenhower Avenue, Alexandria, VA 22314

Physical Address: 490 L'Enfant Plaza SW, Suite 8001, Washington, DC 20024, USA Tel: 202-459-9674, Fax: 202-459-9673, Email: nco@nitrd.gov, Website: https://www.nitrd.gov