DataKitchen

DATAVERSITY

DATAVERSITY Demo Day:

The Data Journey That Leads to Quality and Observability At Scale

November 2023

Current Pain With Data & Analytic Teams

DataKitchen

Data Analytic Projects Fail

60% of projects fail

– Gartner

79% have too many errors

– Eckerson

73% of data practitioners do not trust the their data –

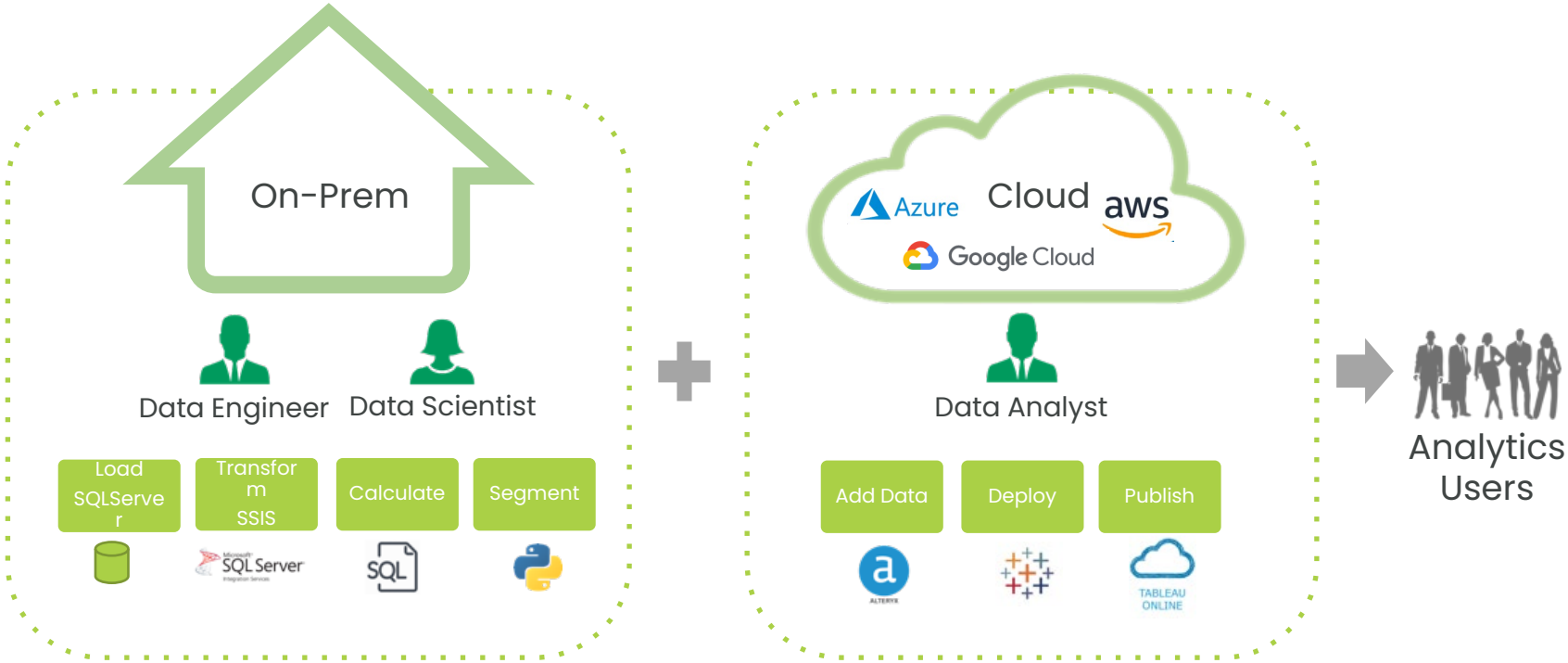
IDC

78% of data teams are stressed & want therapy

-DataKitchen

Challenge: Complexity in Teams, Tools, & Environments

Leads to unreliable Data Journeys, results, and adoption



DataKitchen

Challenge: No End to End Enterprise Observability

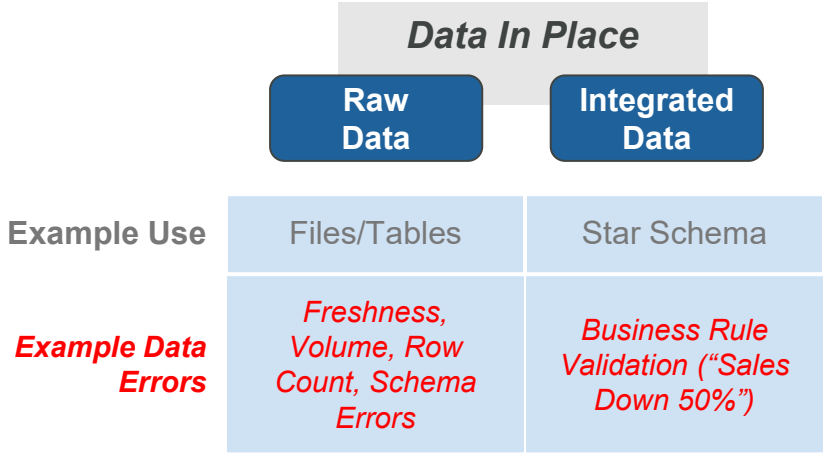
Data Journeys - many data sets, tools, servers and pipelines

- No enterprise-wide visibility of 100's, 1000's of Data Journeys
- No end-to-end quality control
- Hard to diagnose issues
- Errors create distractions that limit development

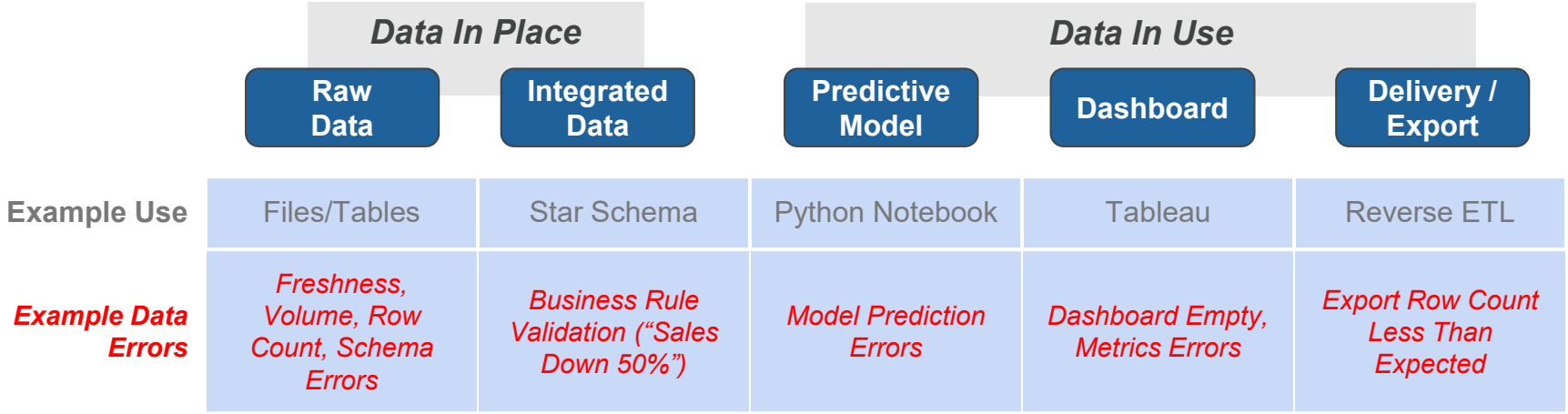
Tool/Component Options	Azure	AWS	GCP	Third Party/others
Big Data ETL/Orchestration	Data Factory	Glue, DataPipeline	DataFlow, Composer (Airflow),	Talend, Informatica. IBM DataStage
Schedule / Workflow	DataBricks	Glue, Kinesis, EMR	PubSub, Composer (Airflow), Cloud DataFusion	Airflow, Streamsets, FiveTran
Data Science	Python, Tensorflow, DataBricks	Python, Sagemaker	Python, AI Platform, Tensorflow	DataRobot, IBM Watson
DevOps	Azure DevOps	Code Deploy, Code Pipeline	CloudBuild	DBT, Delphix, Puppet
Storage	ADLS	S3	GCS	IBM Cloud Object Storage
Databases	Cosmos, SQL Server, Synapse	Redshift, Aurora, DocumentDB	Big Query, Postgres, Big Table	Snowflake, Teradata, Vertica
Analytic Tools	PowerBI	QuickSight	Looker, DataLab, DataStudio	Qlik,Tableau Thoughtspot, Cognos
Infrastructure automation	Terraform	Cloud Formation, Chef, Puppet	Terraform, Chef, Procurement Manager	Chef, Ansible
Data Catalog	Azure Data Catalog	Glue, LakeFormation	Google Data Catalog	Collibra, Watson Catalog
Other Data Tools	Master Data Management; Self Service Data Prep; Code Repositories; Cloud Data Ingestion Tools; Large Language Models, …			

DataKitchen

Multiple Problem Locations In Data Analytics Production



Multiple Problem Locations In Data Analytics Production



DataKitchen

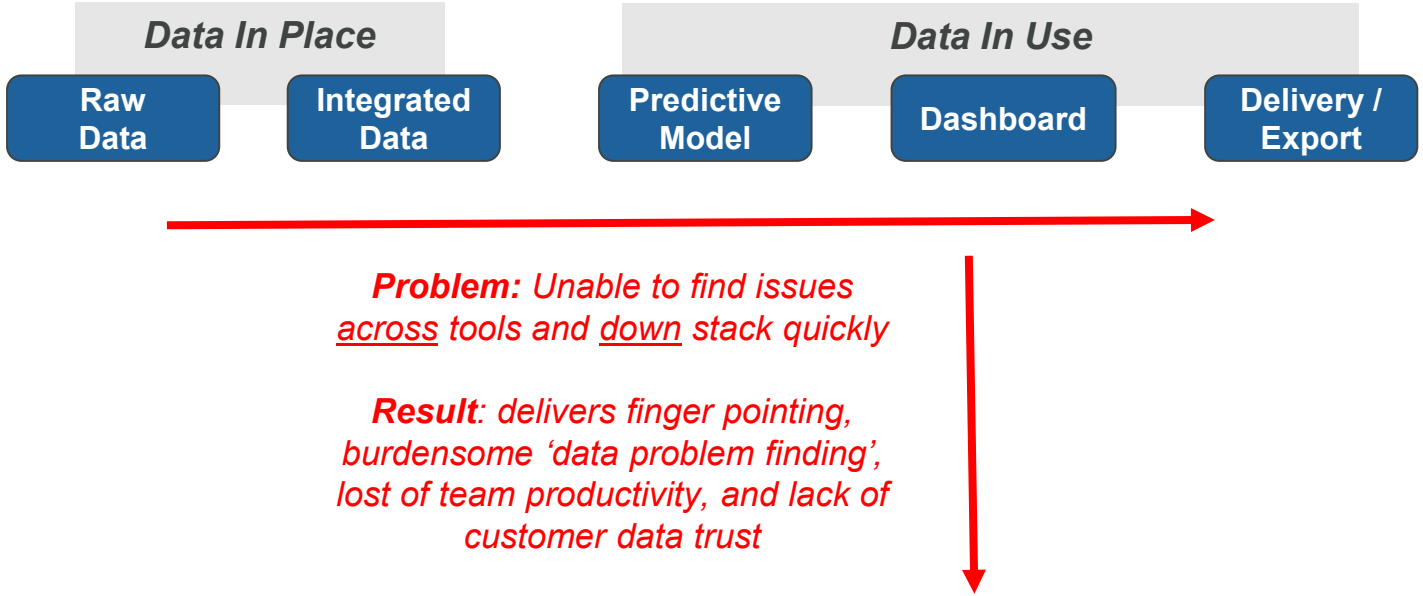
‘Across and Down’ Problems in Data Analytics Production

	Data In Place		Data In Use		
	Raw Data	**Integrated Data**	**Predictive Model**	**Dashboard**	**Delivery / Export**
Example Use	Files/Tables	Star Schema	Python Notebook	Tableau	Reverse ETL
Example Data Errors	*Freshness, Volume, Row Count, Schema Errors*	*Business Rule Validation (“Sales Down 50%”)*	*Model Prediction Errors*	*Dashboard Empty, Metrics Errors*	*Export Row Count Less Than Expected*
Data & Analytic Specific Process Errors	*Run Start / Stop Failures, Schedule Errors, Order of Operations Error, Usage Issues*				

‘Across and Down’ Problems in Data Analytics Production

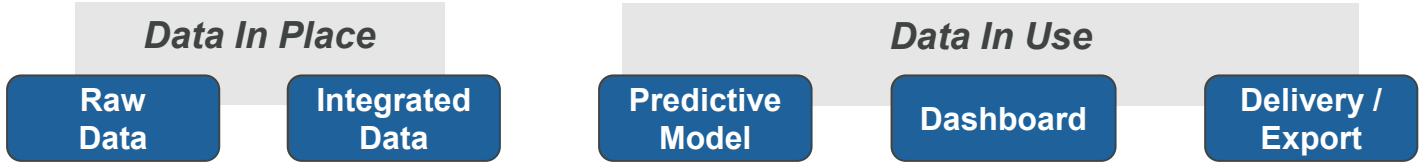
	Data In Place		Data In Use		
	Raw Data	**Integrated Data**	**Predictive Model**	**Dashboard**	**Delivery / Export**
Example Use	Files/Tables	Star Schema	Python Notebook	Tableau	Reverse ETL
Example Data Errors	*Freshness, Volume, Row Count, Schema Errors*	*Business Rule Validation (“Sales Down 50%”)*	*Model Prediction Errors*	*Dashboard Empty, Metrics Errors*	*Export Row Count Less Than Expected*
Data & Analytic Specific Process Errors	*Run Start / Stop Failures, Schedule Errors, Order of Operations Error, Usage Issues*				
Basic IT Level Monitoring Errors	*Log File Errors, Metric Issues (CPU, Disk ..), Cost Issues*				

Correlation Of Errors in Data Analytics Production



DataKitchen

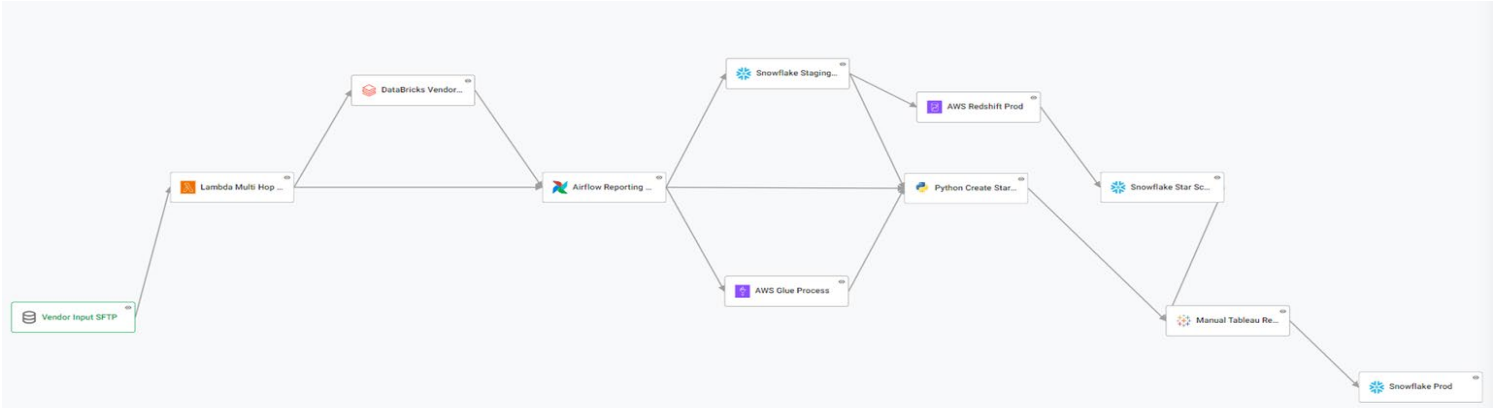
DataKitchen: Observes Data Journeys



Resulting Data Errors

Data & Analytic Specific Process Errors

Basic IT Level Monitoring Errors

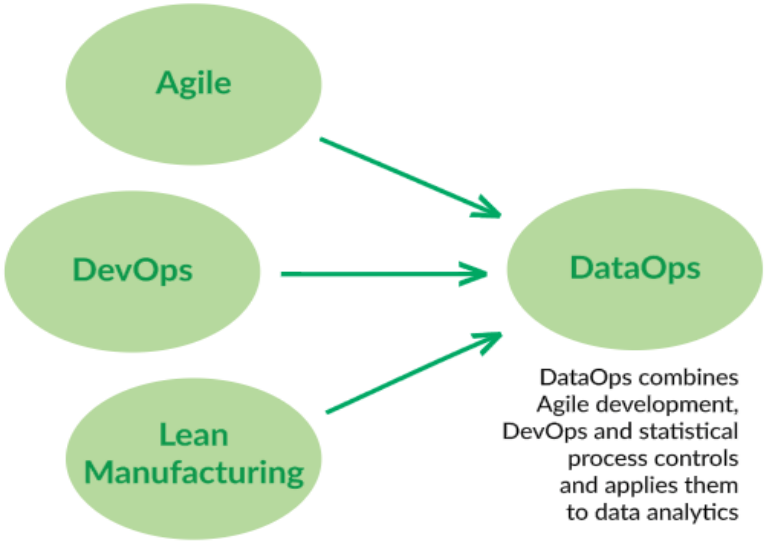


DataKitchen

DataKitchen Data Observability

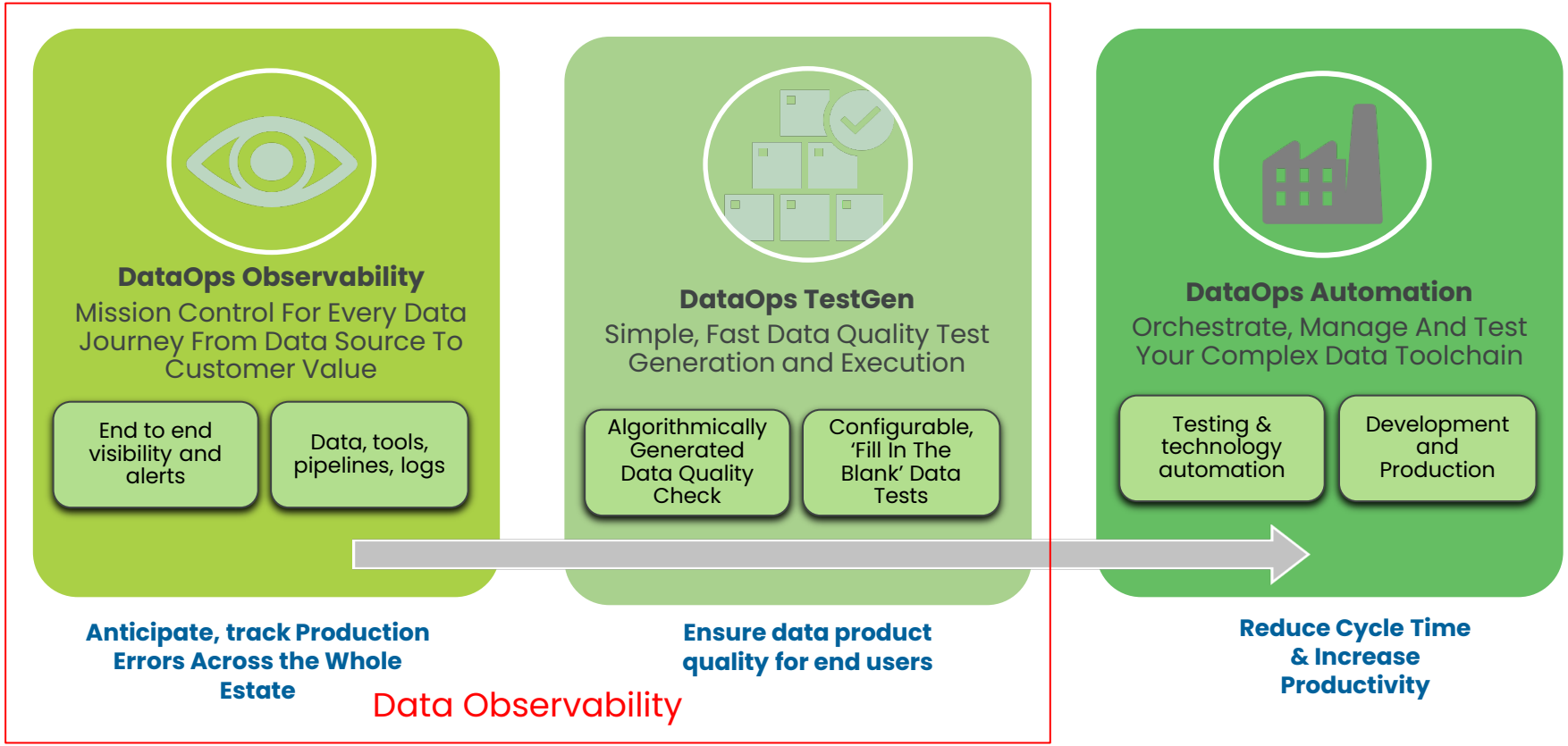
Benefits:

- **Visibility of events** across all journeys:
  - **Data quality test results - TestGen**
  - Infrastructure logs/metrics ***and:***
  - Order of operations
  - Technology/tool status
  - End to end SLA
- **Rapid time to value, any tool chain**
  - No changes to existing pipelines (DataKitchen agents)

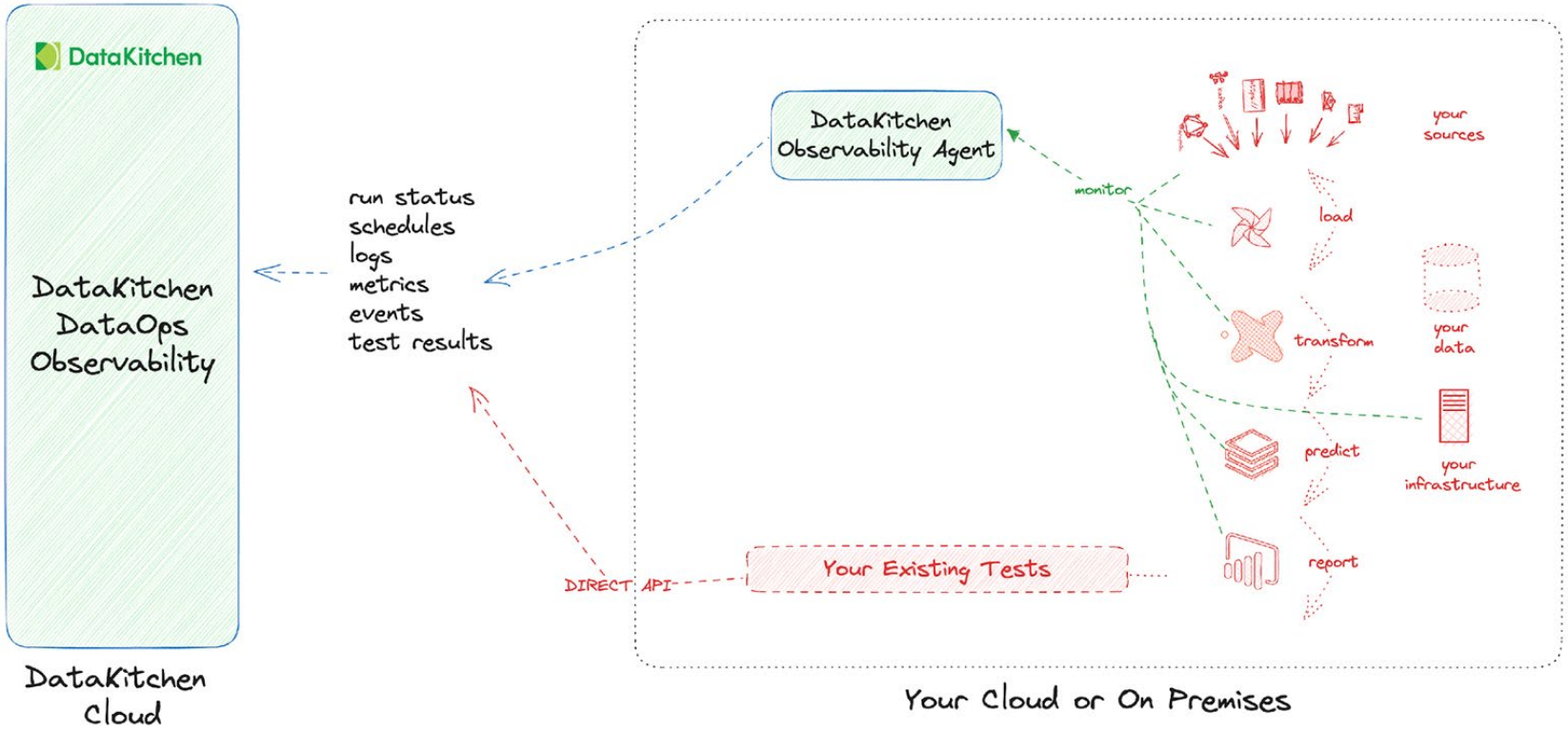


DataKitchen

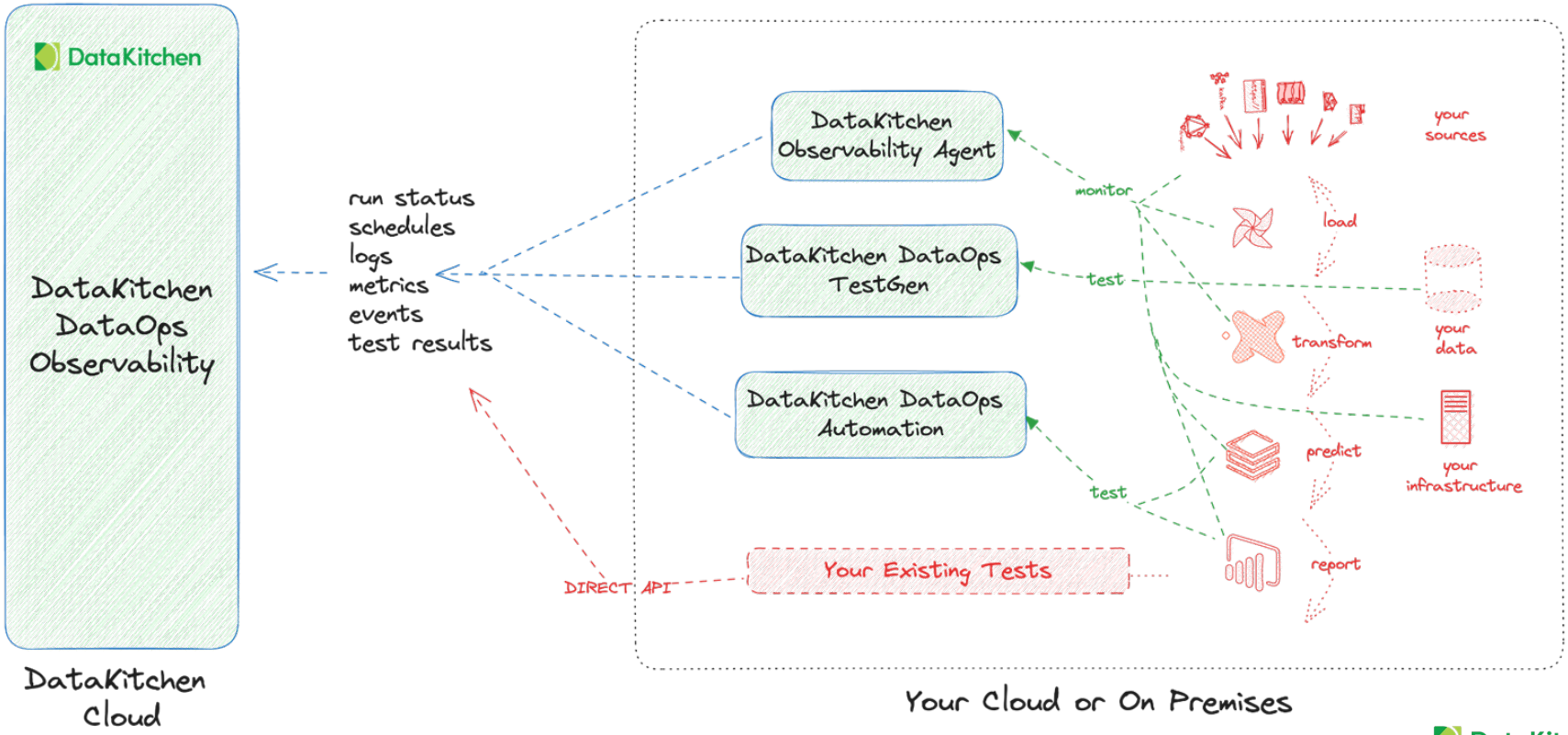
DataKitchen Software Products



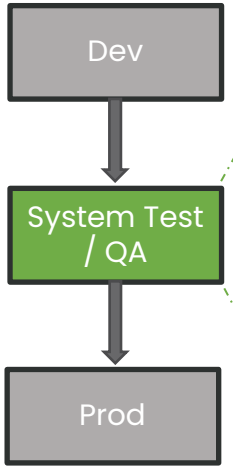
DataKitchen Product Architecture: DataOps Observability and Agents



DataKitchen Product Architecture: DataOps Observability, Test, and Automation

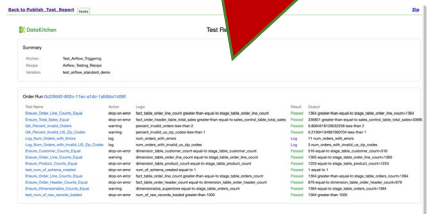


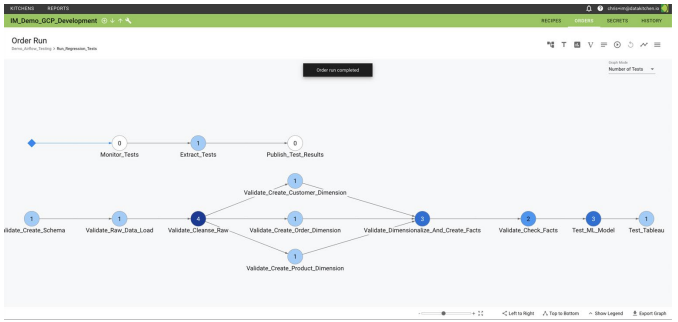
Development Regression/Impact Testing



A CI /CD Process (Jenkins)

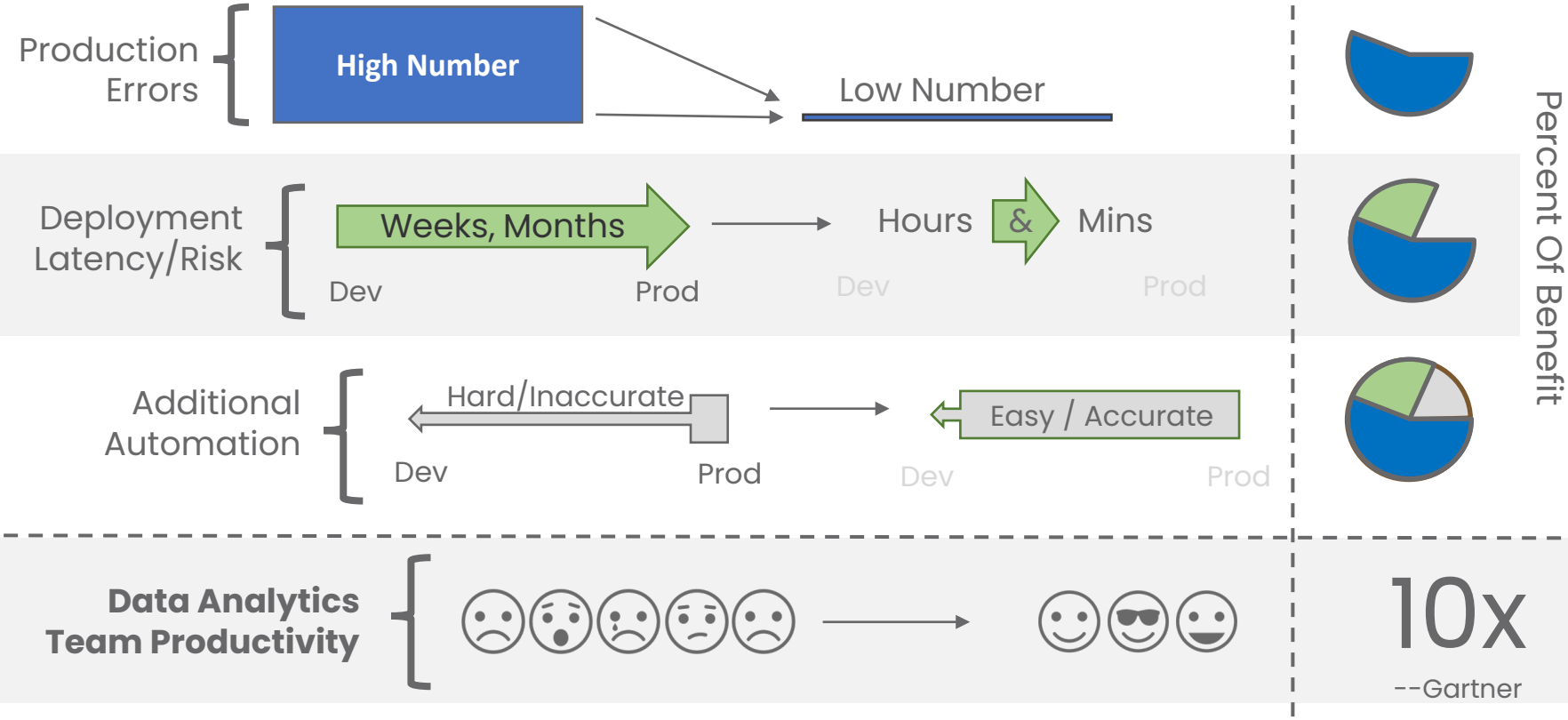
See the end-to-end view



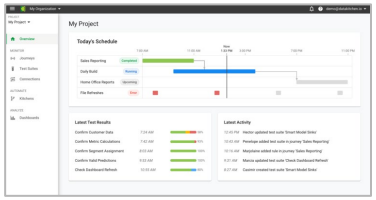


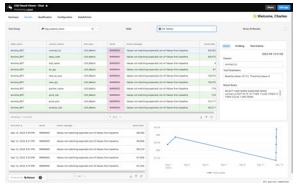
DataKitchen: Faster, Better & Happier

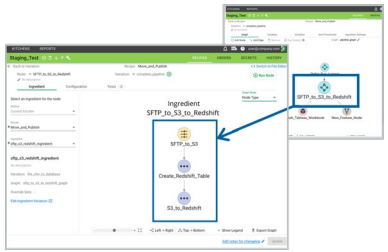
DataKitchen



DataKitchen Software







DataOps Observability:

- **End to End Data Journey Observability** across all tools, data, and infrastructure
- **Complete Toolchain Production Monitoring and Alerting:** Ensure speed & quality of delivery through key metrics and relevant alerts
- **'Mission Control' Dashboards** and historical analytics

DataOps TestGen:

- **Simple, Fast Data Quality Test Generation and Execution**
- **Data Profiling:** database scanning, profiling, and identification of 'bad data.'
- **53 Unique Data Test Types:** algorithmically generated and user-configurable business rule-based configurable run-time data tests and execution

DataOps Automation:

- **Automated Data and Tools Testing:** custom data test development in many languages and APIs
- **Test Development Environments**: Automate the creation & management of environments to speed new feature delivery; secure vault;
- **DataOps Automation**: Common collaboration system (Kitchens) across roles & teams. Meta-Orchestration: design & execute on multi-tool system