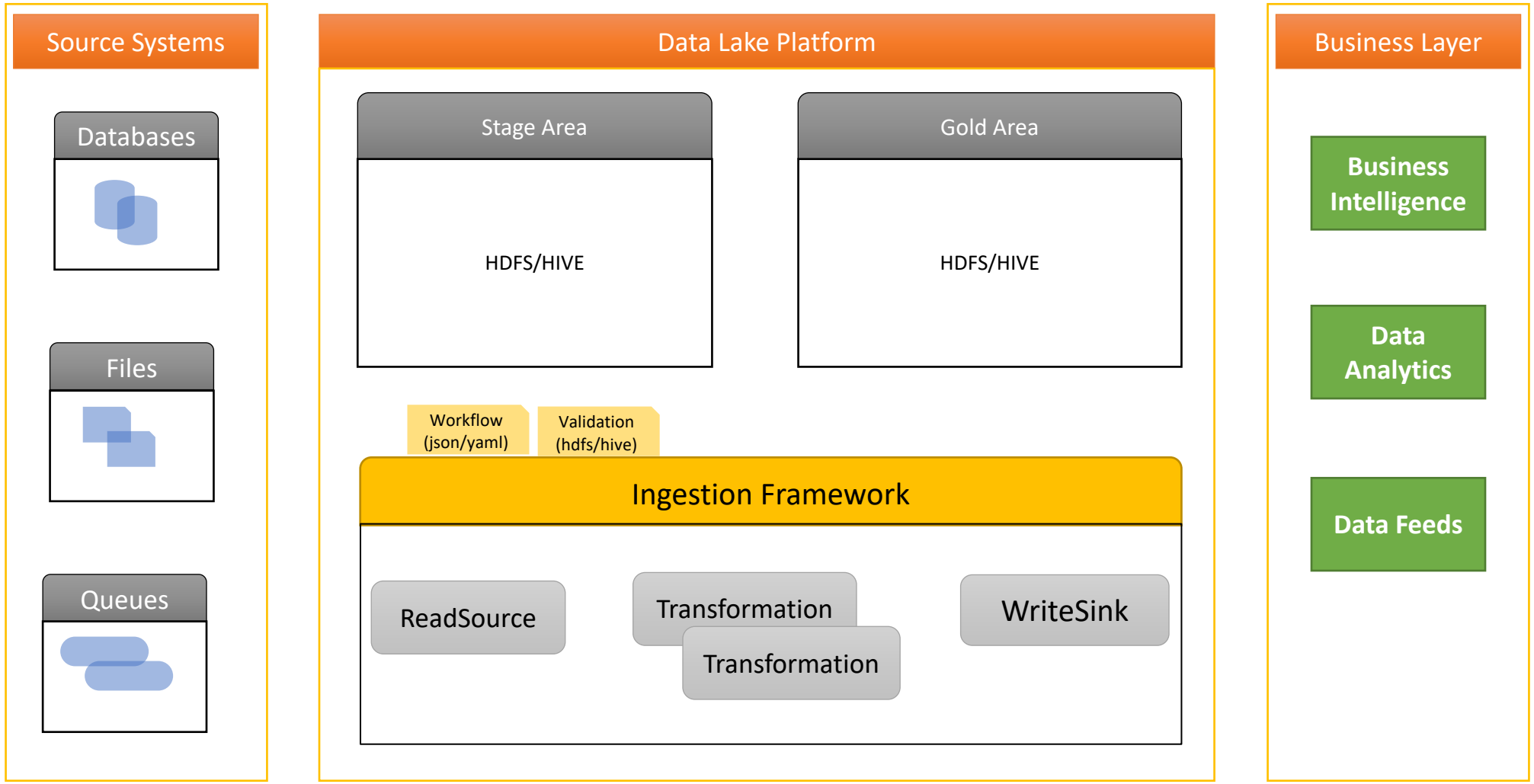
Ingestion Spark Framework

DDA

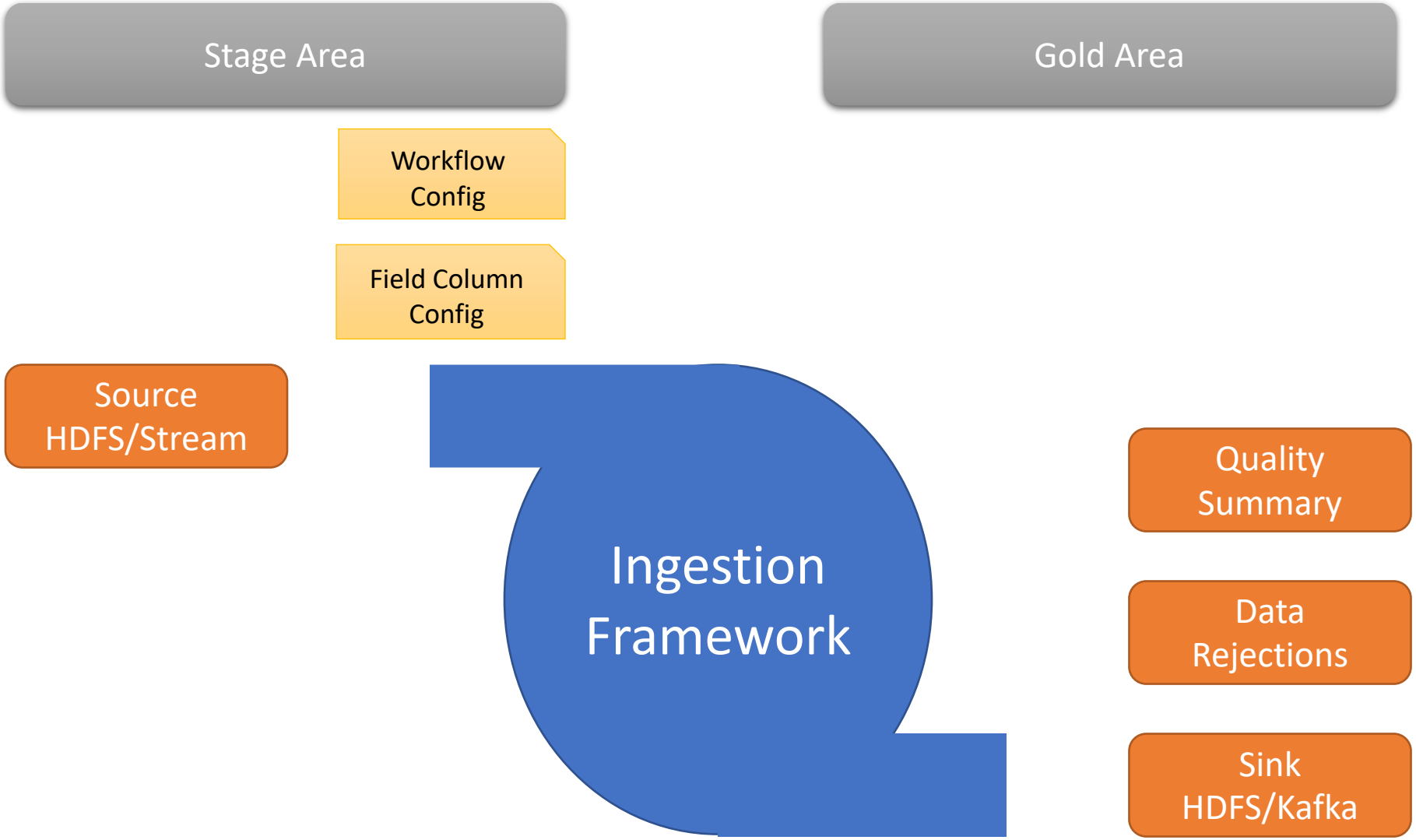
Features

- Drive Ingestion process using Spark
- Define Ingestion Workflow with config files (json/yaml)
- Handle variety source sink formats
- Support transformations
- Share attributes lineage between workflow steps
- Log ingestion process with timestamps
- Field level validation configuration
- Record Rejections

Architecture



Ingestion Flow

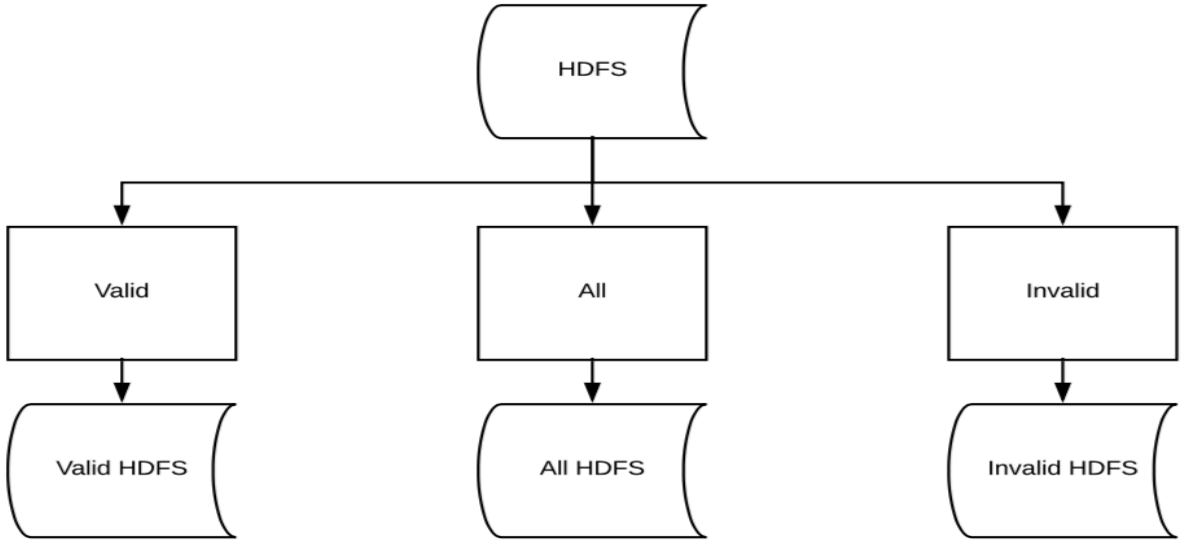


Workflow Definition

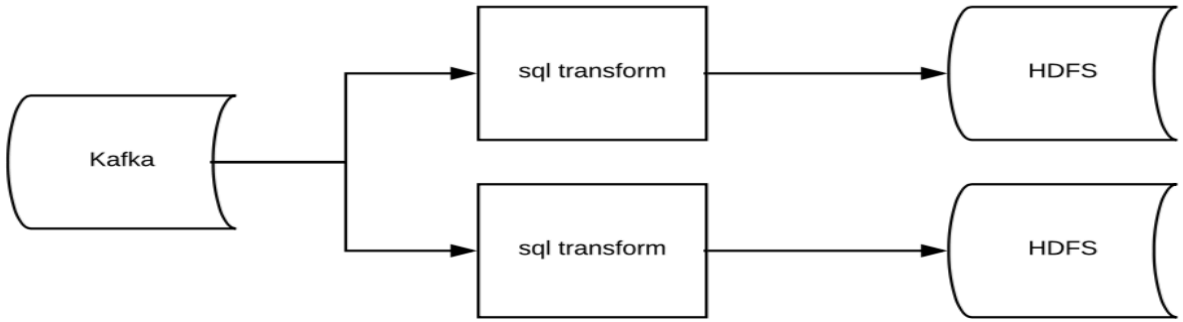
```
---
process: FileFlow
mode: batch
steps:
  - name: readfile
    model: source
    format: csv
    label: filestream
    options:
      header: 'true'
      path: input/sales
    attributes:
      logStatus: 'true'
      statusComment: process started
      statusPath: output/status
      schemaCommand: com.drake.schema.DefaultSchemaBuilder
      schemaPath: schemas/sales.sch
    post:
      category: sh
      name: get_file_info.sh
      path: scripts
```

```
- name: transinvalidfile
    model: transformation
    from: filestream
    command: com.drake.editor.InValidConfigBuilder
    label: invalidstream
  - name: sinkinvalidfile
    model: sink
    from: invalidstream
    format: csv
    options:
      path: output/salesinvalid
      checkpointLocation: output/cpinvalid
    attributes:
      trigger: 10 seconds
```

Workflow Definition



name	from	label
readfile		filestream
transfile	filestream	multiplystream
transvalidfile	filestream	validstream
transinvalidfile	filestream	invalidstream
sinkfile	multiplystream	
sinkvalidfile	validstream	
sinkinvalidfile	invalidstream	



name	from	label
readkafka		kafkastream
transkafka01	kafkastream	transkafkastream01
transkafka01	kafkastream	transkafkastream02
sinkconsole01	transkafkastream01	
sinkconsole02	transkafkastream02	

name	from	label
readhive		hivestream
transhive	hivestream	hivepartstream
sinkhive	hivepartstream	

Data Quality Configuration – Data Columns

db_name	table_name	col_name	data_type	data_type_check	null_check	empty_check	dup_check	range_check	range_check_values	const_check	const_check_values
default	sales_stg	transactionId	Long	1	1	1	1	0		0	
default	sales_stg	customerId	Long	1	1	1	1	0		0	
default	sales_stg	itemId	Long	1	1	1	1	0		0	
default	sales_stg	amountPaid	Double	1	1	1	1	0		0	

Data Quality Summary

db_name	table_name	col_name	sum_data_type_check	sum_null_check	sum_dup_chk	sum_empty_check
default	sales_stg	transactionId	0	0	0	0
default	sales_stg	customerId	1	1	12	0
default	sales_stg	itemId	1	1	13	0
default	sales_stg	amountPaid	1	0	10	0

Sales Summation Chart

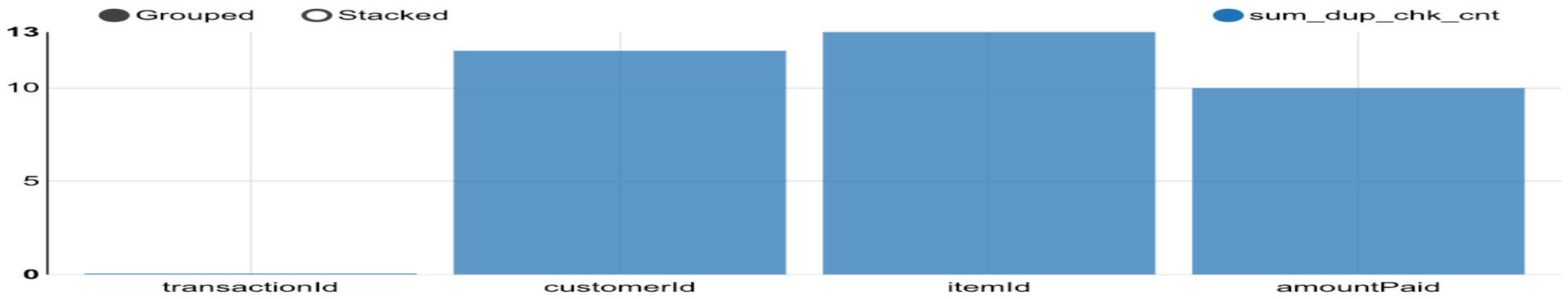
FINISHED

```
%sql
select col_name, sum_dup_chk_cnt
from invalid_col_stats
where src_system='${system=sales,sales|cnesus}' and table_name='${system=sales
    ,sales|cnesus}' and event_dt='${date}'
```

system: sales

date: 2018-12-25

settings



Invalid Rejection Records

transactionId	custormerId	itemId	amountPaid	datatypechkstat	nullchkstat	emptychkstat	Stat
123	1	4	500.00				
124	1	null	500.00	itemId	itemId		DataTypeCheck:itemId NullCheck:itemId
125	null	4	xx	customerId~amountPaid	customerId		DataTypeCheck:customerId~amountPaid NullCheck:customerId

Workflow						
						A name given to workflow
				process	stream & batch	process: KafkaConsoleFlow
						Either "batch" which specifies it is batch workflow or "stream" which specifies this is streaming pipeline
				mode	stream & batch	mode: stream
steps						A sequence of steps to defines batch or streaming pipeline which source, sink and transformations
	Source					Defines type source where data is read from
		Kafka				A Kafka Stream source
				name	readkafka stream & batch	Uniquely defined name in the pipeline
				model	source stream & batch	Type of step source/sink/transformation
				format	kafka stream & batch	Type of source
				from		A name to help to build links between steps in the pipeline where it gets the data from
				label	kafkastream stream & batch	A name to help to build links between steps in the pipeline
						key-value pairs passed directly to spark as options attribute
						kafka.bootstrap.servers: localhost:9092
						subscribe: census
			options		stream & batch	startingOffsets: earliest
			attributes		stream & batch	key-value pairs will direct the spark logic
				logStatus	false	Log the process started status into status table/file
						To convert Kafka incoming meesage
				kafkaSelectExpr		kafkaSelectExpr: CAST(value AS STRING) as json
	Transformation					
				name	transkafka stream & batch	Uniquely defined name in the pipeline
				model	transformation stream & batch	Type of step source/sink/transformation
				from	kafkastream stream & batch	A name to help to build links between steps in the pipeline where it gets the data from
						A handler class to take action using the attributes/conversions/incl/post configurations
				command	stream & batch	com.drake.editor.DefaultEditorBuilder
				label	transkafkastream stream & batch	A name to help to build links between steps in the pipeline
						key-value pairs will be included in the out going dataframe
						attributes:
						category: schema
						fromJsonColumn: json
						fromJsonAlias: json
			attributes		stream & batch	schemaCommand: com.drake.schema.DefaultSchemaBuilder
						key-value pairs to instruct to generate new dataframe
						conversions:
						- seq: '1'
						category: sql
						currentTempView: census_02_1
						sql: select * from census_02_1 where age > 50
						- seq: '2'
						category: sql
						currentTempView: census_02_2
			conversions		stream & batch	sql: select * from census_02_2 where age > 90
						key-value pairs will be included in the out going dataframe
						process_dt: "$readfile#fileDt"
			include		stream & batch	process_hr: "$readfile#fileHr"
						key-value pairs will be included in the out going dataframe
						category: sql
						currentTempView: census_01
			post		stream & batch	sql: select * from census_01 where age <= 50
	Sink					
				name	sinkkafka stream & batch	Uniquely defined name in the pipeline
				model	sink stream & batch	Type of step source/sink/transformation
				from	transkafkastream stream & batch	A name to help to build links between steps in the pipeline where it gets the data from
				format	orc stream & batch	Type of data format to be written as to sink
						key-value pairs passed directly to spark as options attribute
						numRows: 5
						truncate: 'false'
						path: output/census02
			options		stream & batch	checkpointLocation: output/cpcensus02
						key-value pairs will direct the spark logic
			attributes		stream & batch	trigger: 10 seconds