TensorFlow
TensorFlow Extended (TFX)
适用于生产环境的端到端机器学习平台

机器学习不仅仅是模型训练...

模型训练

... 生产环境需要更多的工作

配置

数据采集

数据验证

系统监测

模型训练

分析工具

模型部署

进程管理

资源管理

特征提取

TensorFlow Extended (TFX)

端到端的机器学习平台

用户界面：支持进程管理，监测，调试，数据/模型/结果可视化

共享配置与业务调度器

数据提取 | 数据分析与验证 | 数据/特征转换 | **模型训练** | 模型验证 | 模型部署 | 生产日志

存储回收/权限管理

外部存储

谷歌内部用户

nest

DeepMind

verily

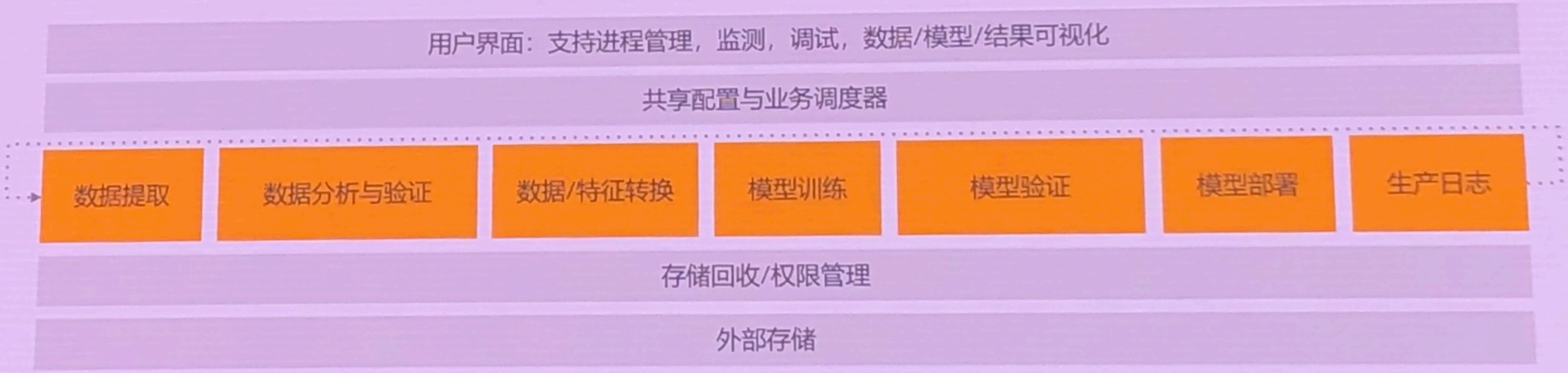
LOON

TensorFlow
分享简要
1. 流水线（Pipeline）
2. 组件（Component）
3. 调度器（Orchestration）和元数据（Metadata）
4. 完整的实例介绍

TensorFlow

什么是流水线（TFX Pipeline）

流水线：一系列组件


用户界面：支持进程管理，监测，调试，数据/模型/结果可视化
共享配置与业务调度器
数据提取
数据分析与验证
数据/特征转换
模型训练
模型验证
模型部署
生产日志
存储回收/权限管理
外部存储

横向分层：协调组件关系

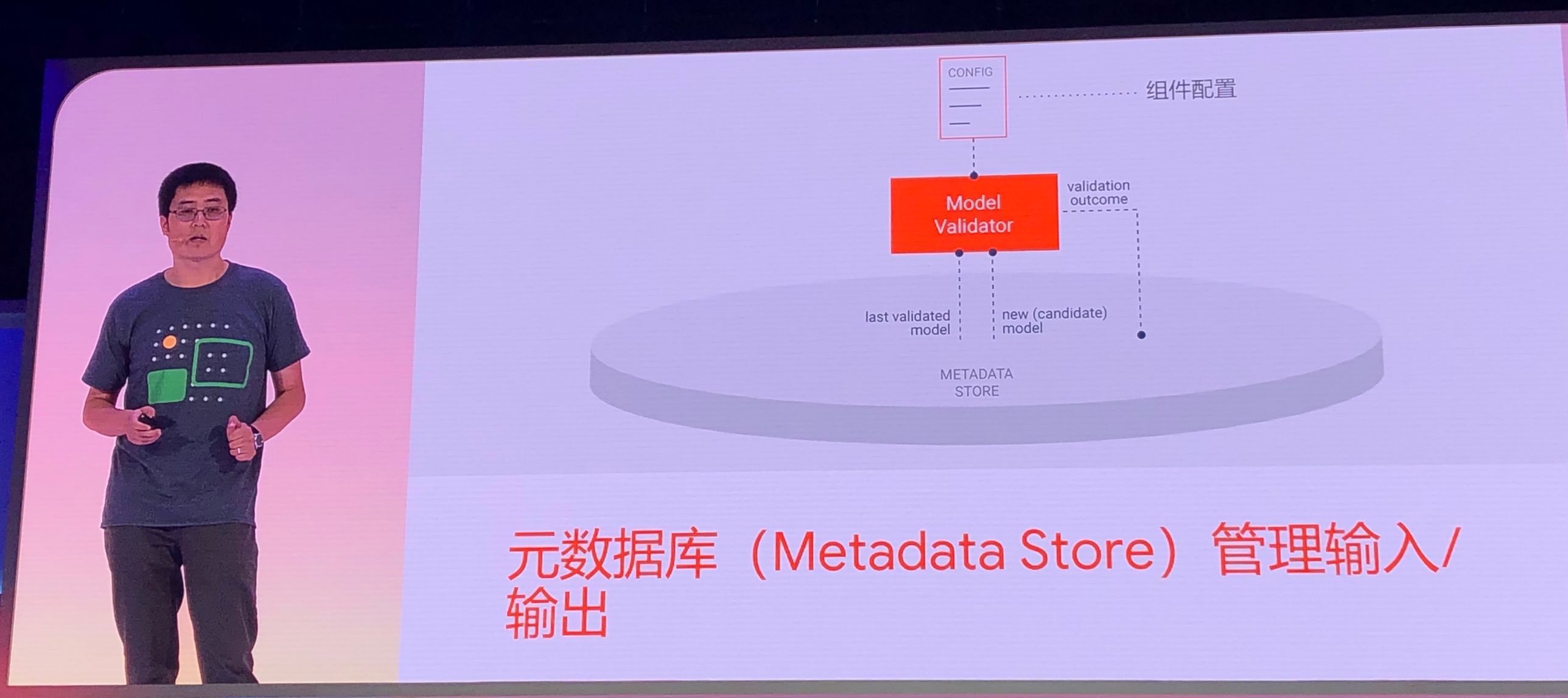
用户界面：支持进程管理，监测，调试，数据/模型/结果可视化
共享配置与业务调度器
数据提取
数据分析与验证
数据/特征转换
模型训练
模型验证
模型部署
生产日志
存储回收/权限管理
外部存储

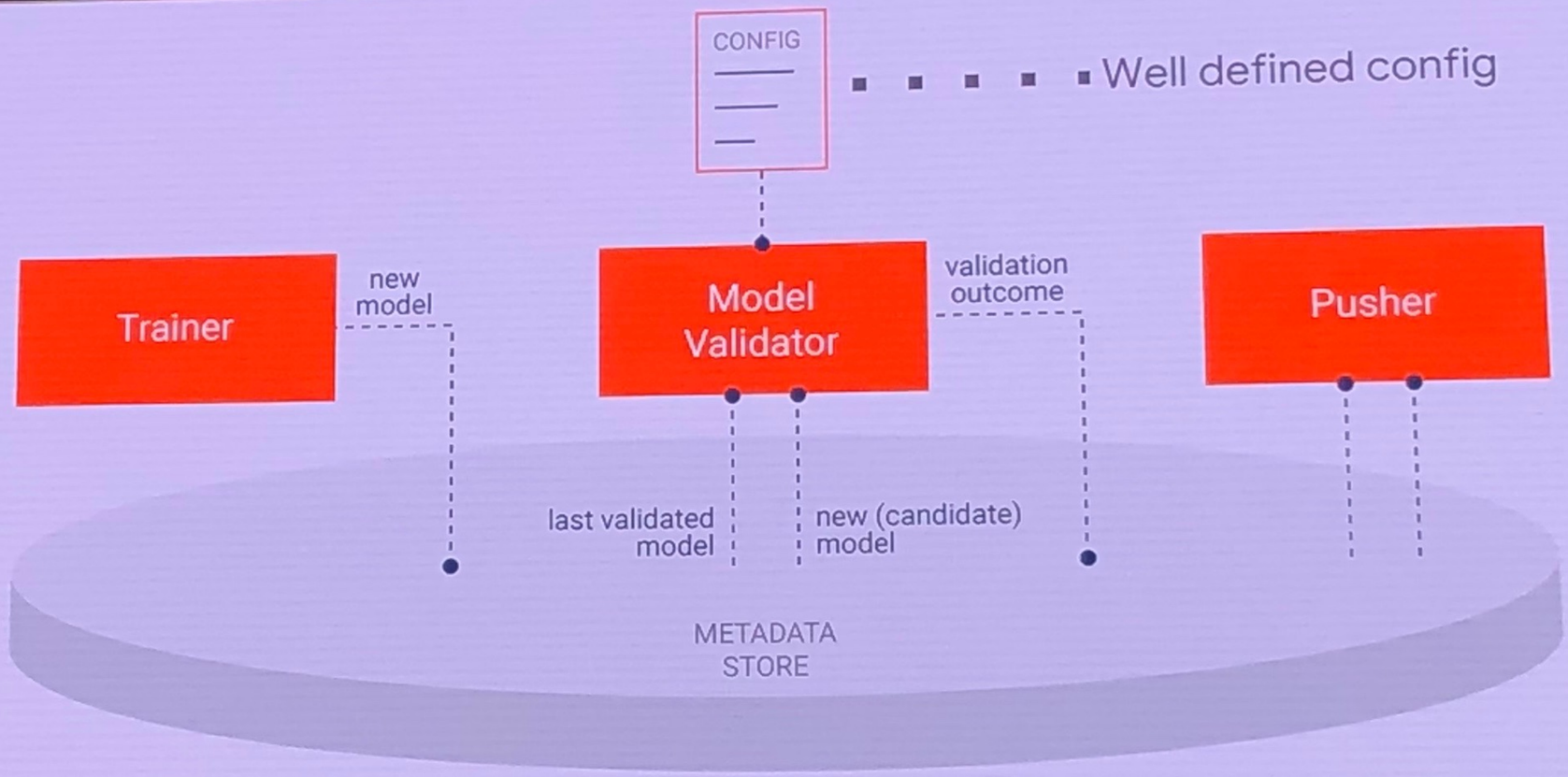
Model
Validator

TFX组件

Model
Validator
validation
outcome
last validated
model
new (candidate)
model
完整的输入/输出定义

CONFIG
组件配置
Model Validator
validation outcome
last validated model
new (candidate) model
组件配置

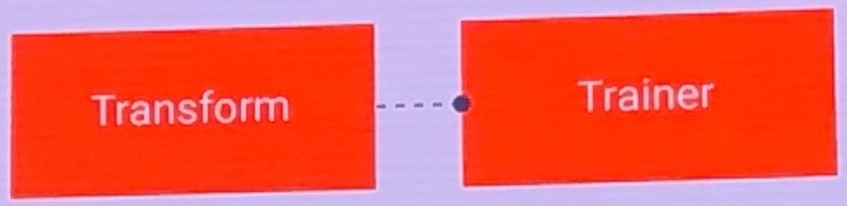

CONFIG
组件配置
Model Validator
validation outcome
last validated model
new (candidate) model
METADATA STORE
元数据库（Metadata Store）管理输入/输出



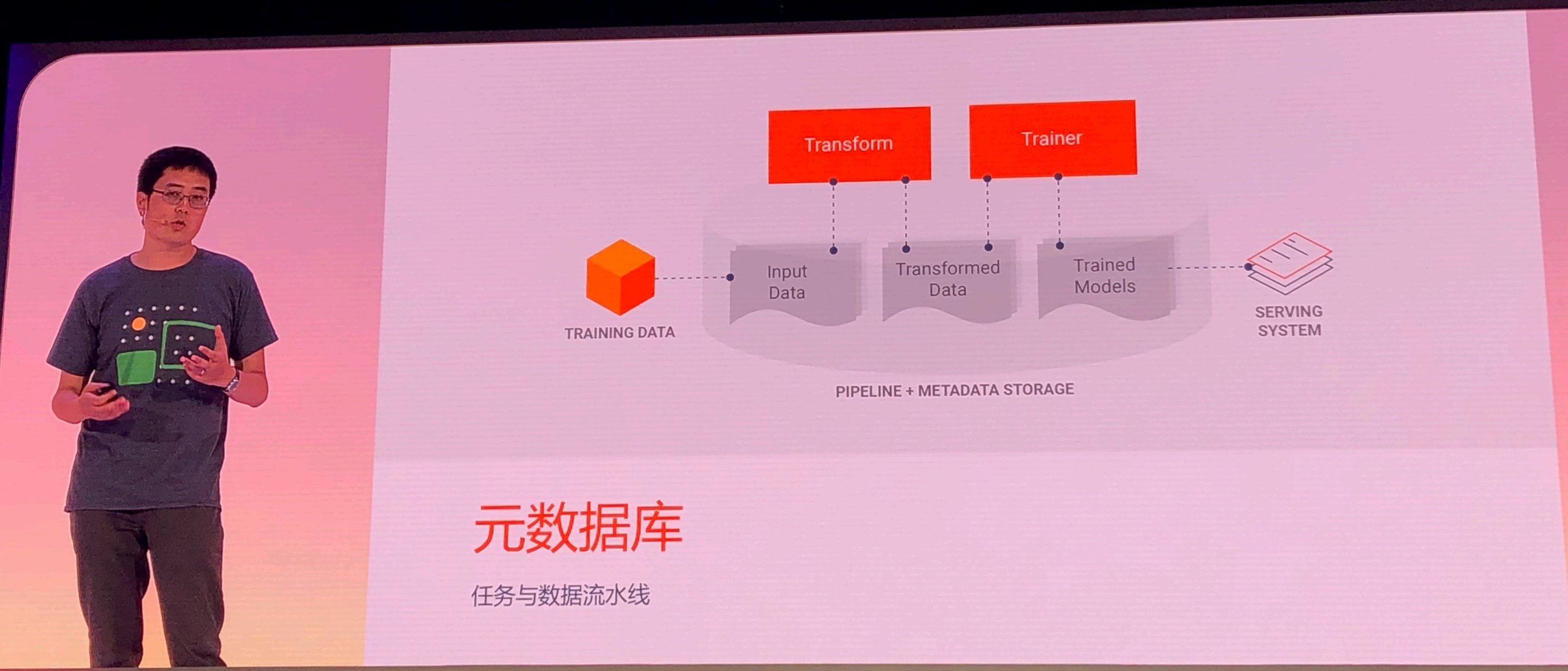
组件间由元数据（Metadata）连接

TensorFlow

调度器（Orchestration）和元数据（Metadata）



任务流水线



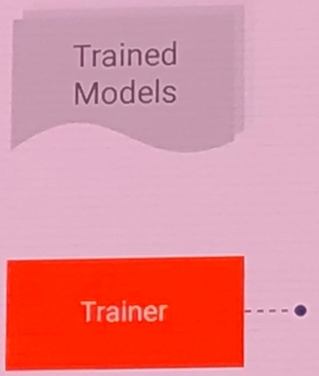
元数据库

任务与数据流水线

Trained
Models

元数据库内容

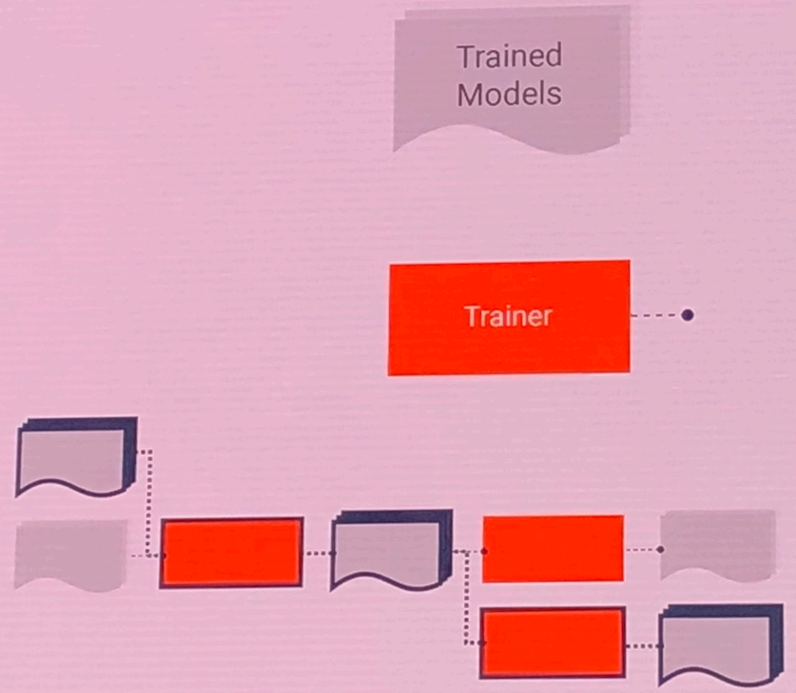
产物（artifact）类型定义及属性



元数据库内容

产物（artifact）类型定义及属性

组件的执行记录

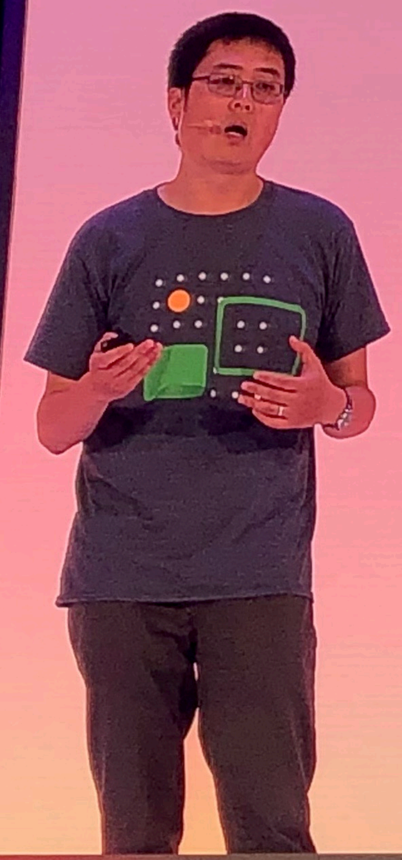


元数据库内容

产物（artifact）类型定义及属性

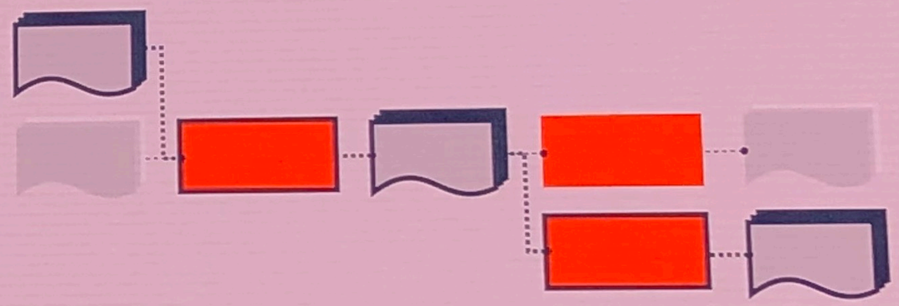
组件的执行记录

溯源(Workflow Provenance)



元数据库功能

回溯模型训练的输入数据

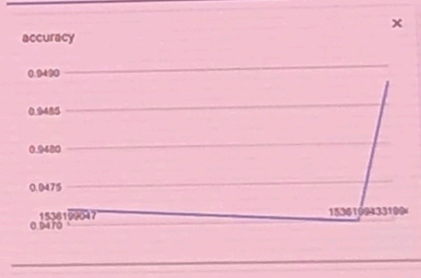


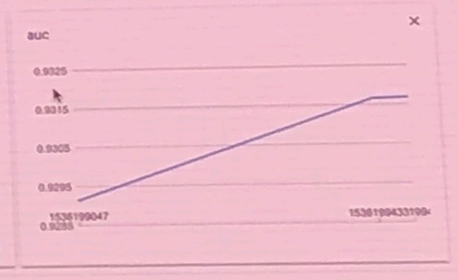
元数据库功能

比较历史版本的模型性能

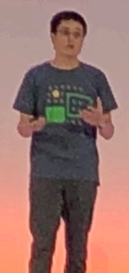
```
eval_results = tfma.make_eval_results([tfma_result_1, tfma_result_2, tfma_result_3],
                                      tfma.constants.MODEL_CENTRIC_MODE)
tfma.view.render_time_series(eval_results, OVERALL_SLICE_SPEC)
```

Add metric series



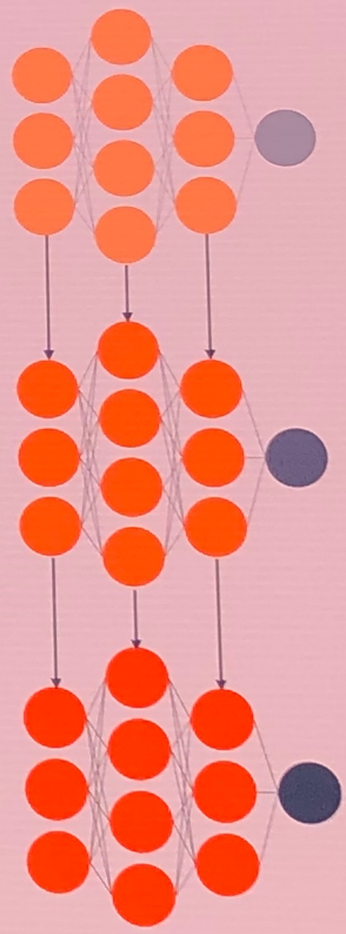


Model	Data	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss	label/mean
1536199479	data.csv	0.94880	0.94220	0.93168	0.98516	0.13980	0.94220
1536199433	data.csv	0.94700	0.94220	0.93165	0.98170	0.13979	0.94220
1536199047	data.csv	0.94720	0.94220	0.92914	0.99480	0.14103	0.94220



元数据库功能

用已有模型版本加速训练

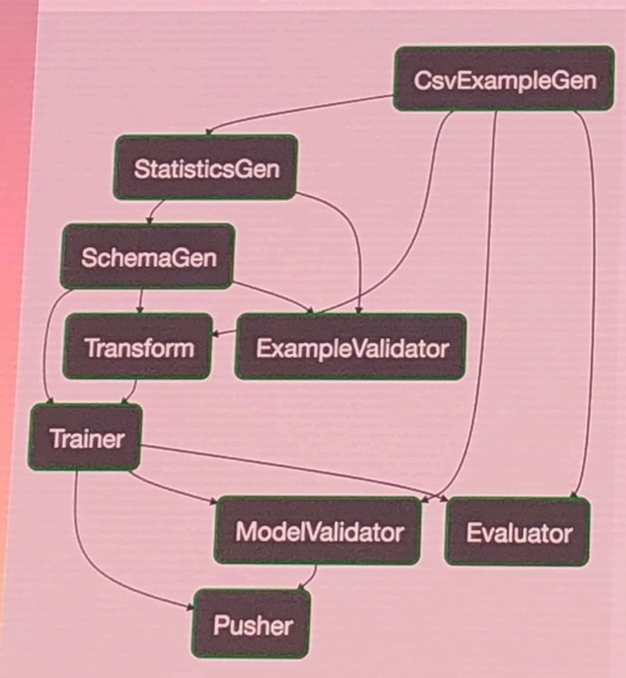


元数据库功能
重用计算结果

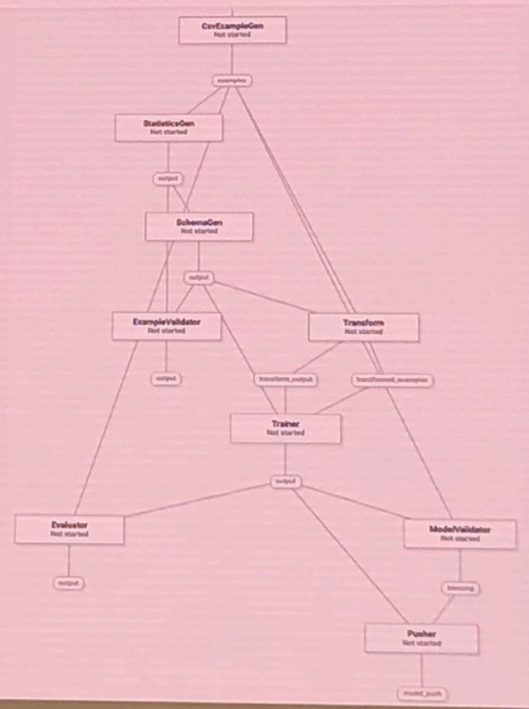
TFX CONFIG
AIRFLOW RUNTIME
KUBEFLOW RUNTIME
OTHER
DRIVER
EXECUTOR
PUBLISHER
DRIVER
EXECUTOR
PUBLISHER
DRIVER
EXECUTOR
PUBLISHER
METADATA STORE
自定义调度器(Orchestrator)
方便的使用不同的作业调度器系统

可移植的调度器系统

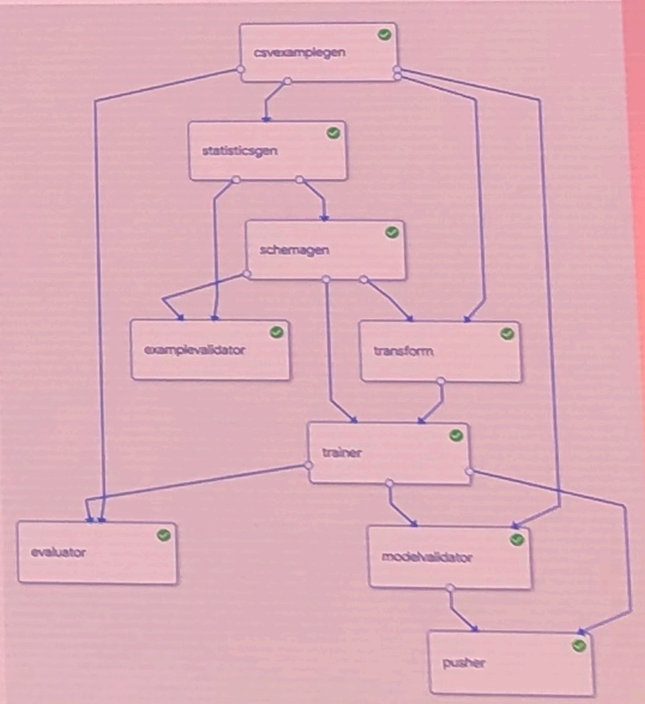
Apache Airflow



Apache Beam



Kubeflow Pipelines



TFX 完整实例

芝加哥出租车数据集

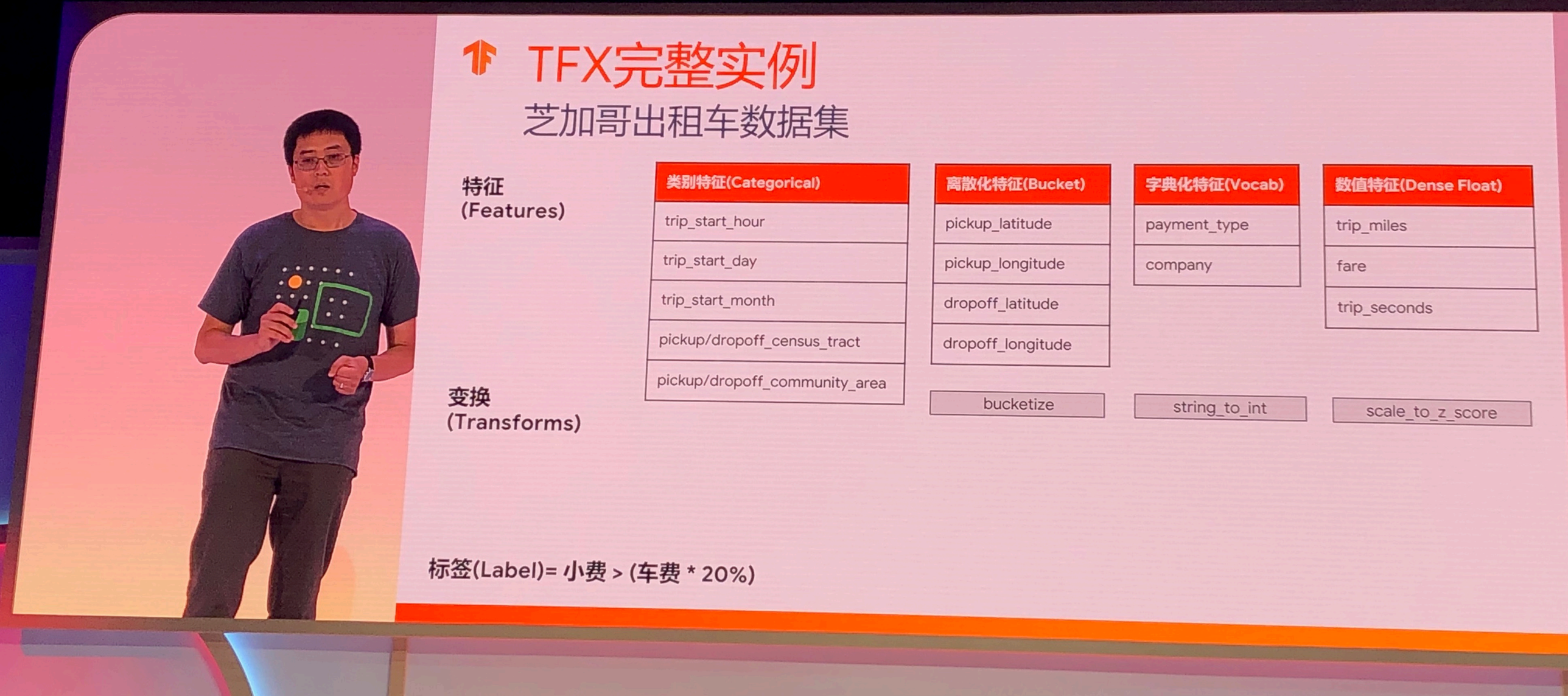
TFX完整实例

芝加哥出租车数据集

特征 (Features)

类别特征(Categorical)	离散化特征(Bucket)	字典化特征(Vocab)	数值特征(Dense Float)
trip_start_hour	pickup_latitude	payment_type	trip_miles
trip_start_day	pickup_longitude	company	fare
trip_start_month	dropoff_latitude		trip_seconds
pickup/dropoff_census_tract	dropoff_longitude		
pickup/dropoff_community_area			

标签(Label)= 小费 > (车费 * 20%)


TFX完整实例
芝加哥出租车数据集
特征
(Features)
类别特征(Categorical)
trip_start_hour
trip_start_day
trip_start_month
pickup/dropoff_census_tract
pickup/dropoff_community_area
离散化特征(Bucket)
pickup_latitude
pickup_longitude
dropoff_latitude
dropoff_longitude
字典化特征(Vocab)
payment_type
company
数值特征(Dense Float)
trip_miles
fare
trip_seconds
变换
(Transforms)
bucketize
string_to_int
scale_to_z_score
标签(Label)= 小费 > (车费 * 20%)

TFX完整实例

芝加哥出租车数据集

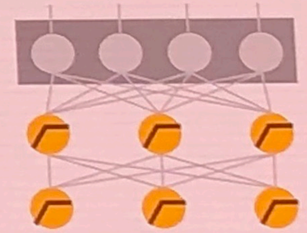
特征 (Features)

类别特征(Categorical)	离散化特征(Bucket)	字典化特征(Vocab)	数值特征(Dense Float)
trip_start_hour	pickup_latitude	payment_type	trip_miles
trip_start_day	pickup_longitude	company	fare
trip_start_month	dropoff_latitude		trip_seconds
pickup/dropoff_census_tract	dropoff_longitude		
pickup/dropoff_community_area			

变换 (Transforms)

bucketize | string_to_int | scale_to_z_score

模型 (Wide+Deep)



标签(Label)= 小费 > (车费 * 20%)

TFX完整实例

芝加哥出租车数据集

特征
(Features)

类别特征(Categorical)
trip_start_hour
trip_start_day
trip_start_month
pickup/dropoff_census_tract
pickup/dropoff_community_area

离散化特征(Bucket)
pickup_latitude
pickup_longitude
dropoff_latitude
dropoff_longitude

字典化特征(Vocab)
payment_type
company

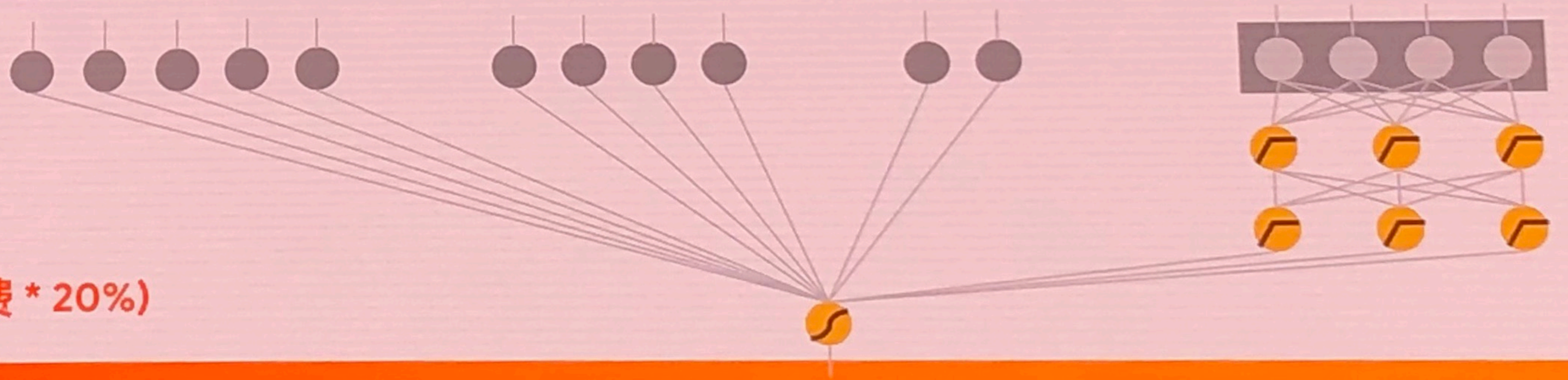
数值特征(Dense Float)
trip_miles
fare
trip_seconds

变换
(Transforms)
模型
(Wide+Deep)

bucketize

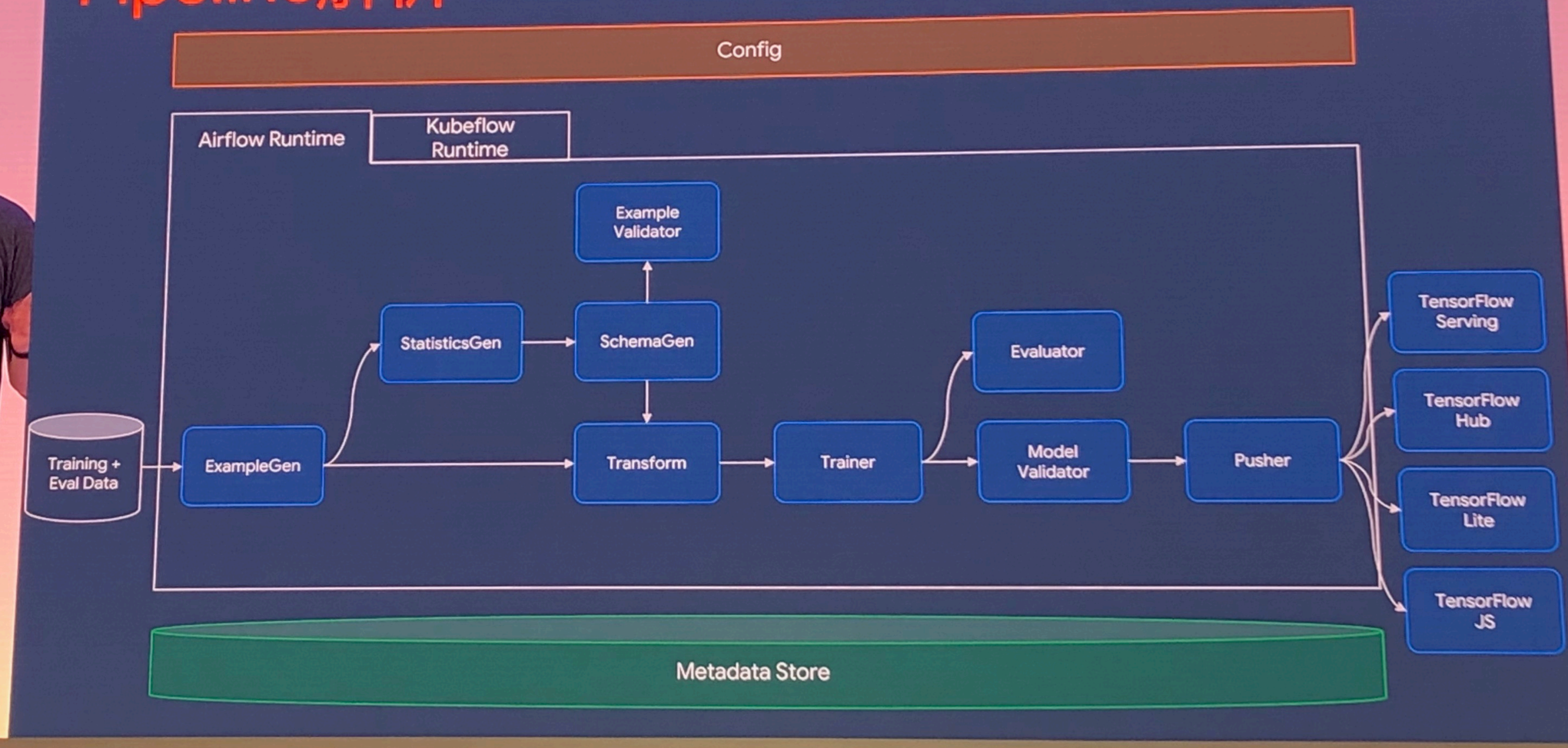
string_to_int

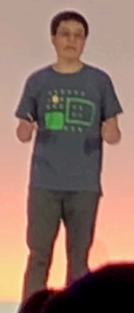
scale_to_z_score



标签(Label)= 小费 > (车费 * 20%)

Pipeline解析

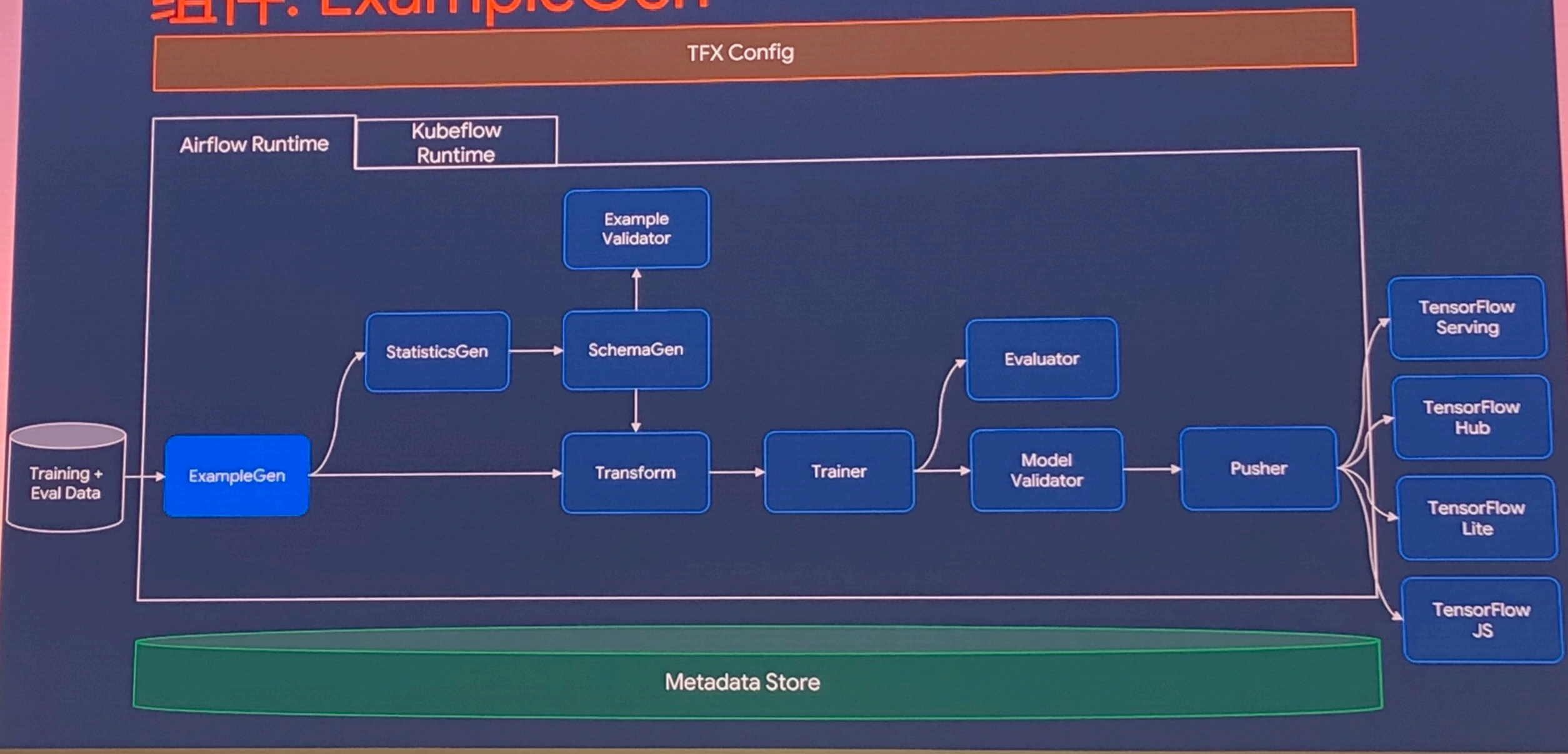

Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Evaluator
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Training + Eval Data
ExampleGen
Transform
Trainer
Model Validator
Pusher
Metadata Store



TensorFlow

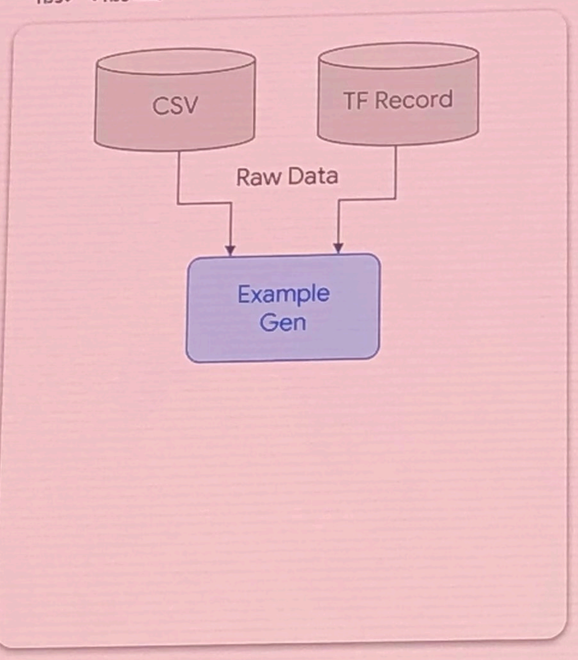
数据提取(Data Ingestion)

组件: ExampleGen



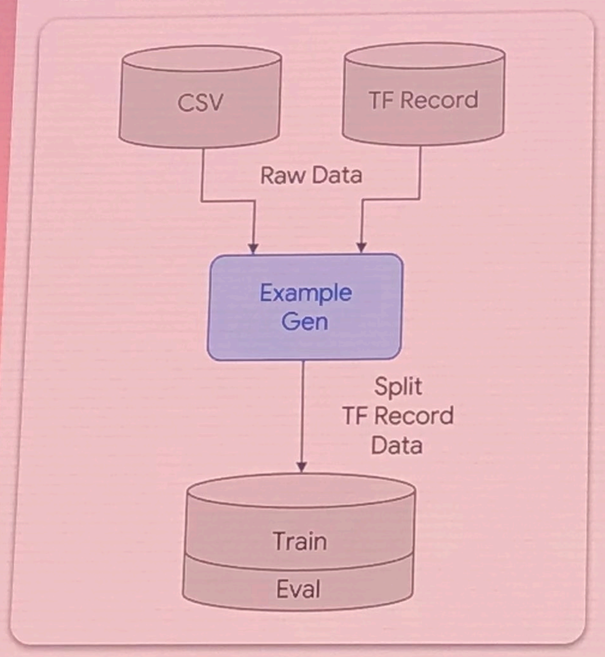
组件: ExampleGen

输入输出



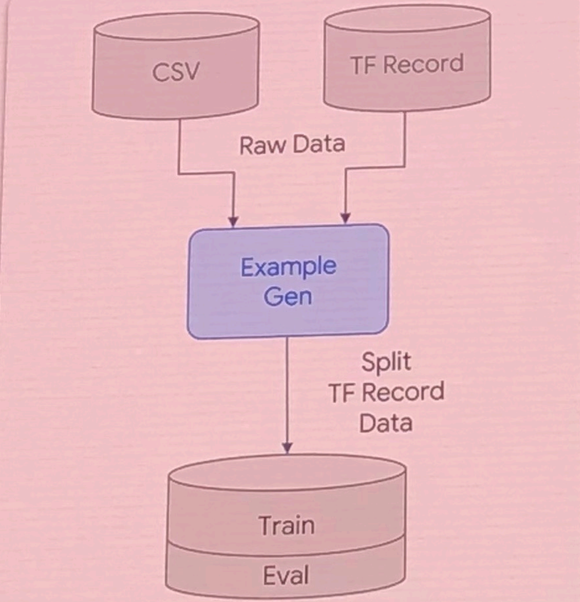
组件: ExampleGen

输入输出



组件: ExampleGen

输入输出

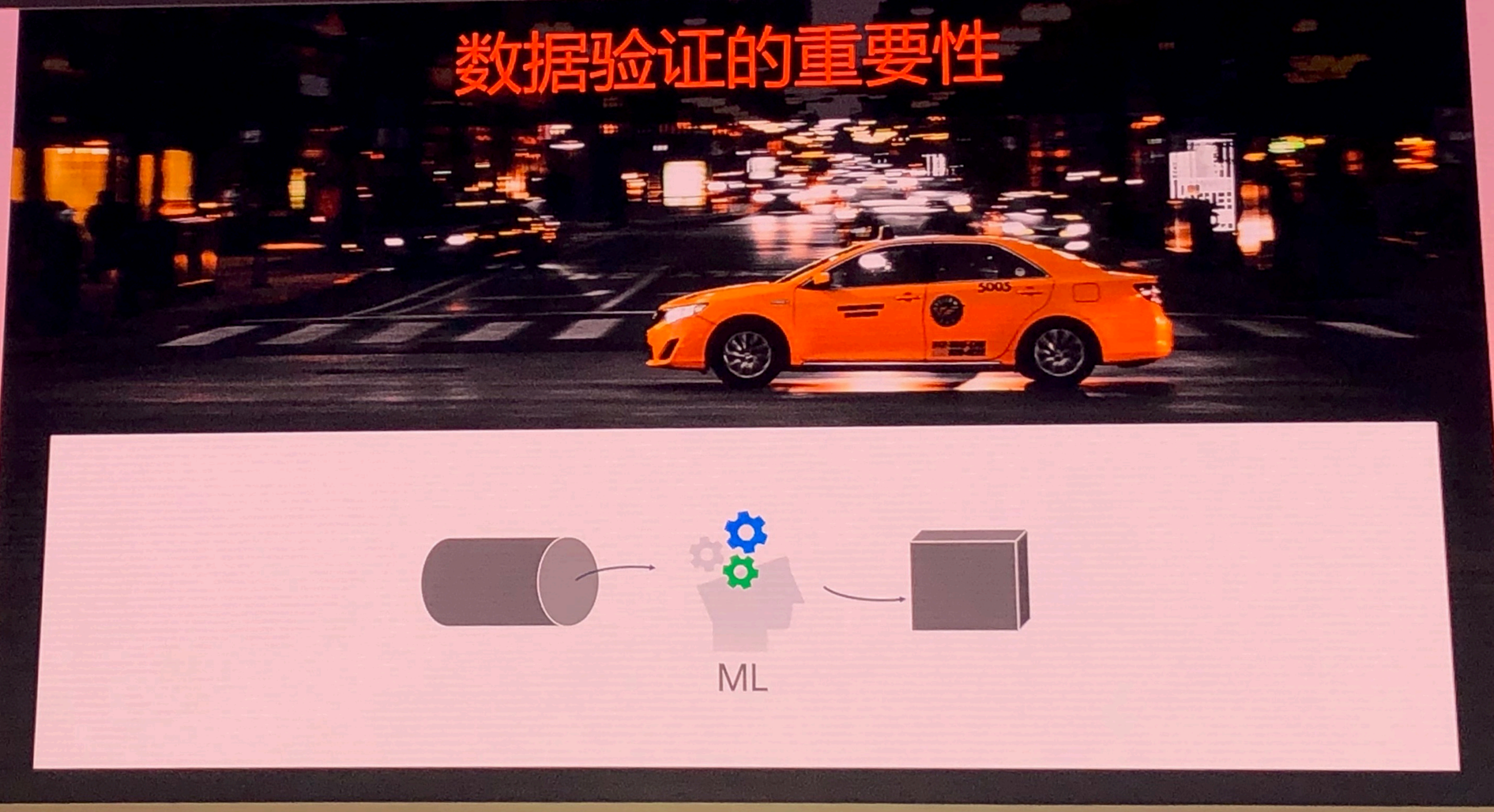


配置

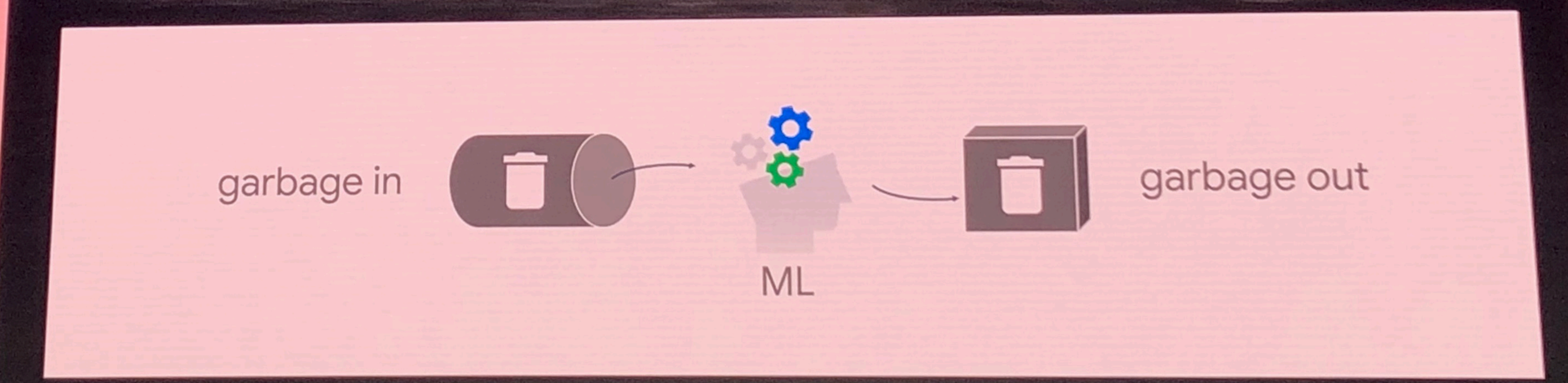
```
examples = csv_input(os.path.join(data_root, 'simple'))
example_gen = CsvExampleGen(input_base=examples)
```

TensorFlow
数据分析和验证(Data Analysis and Validation)

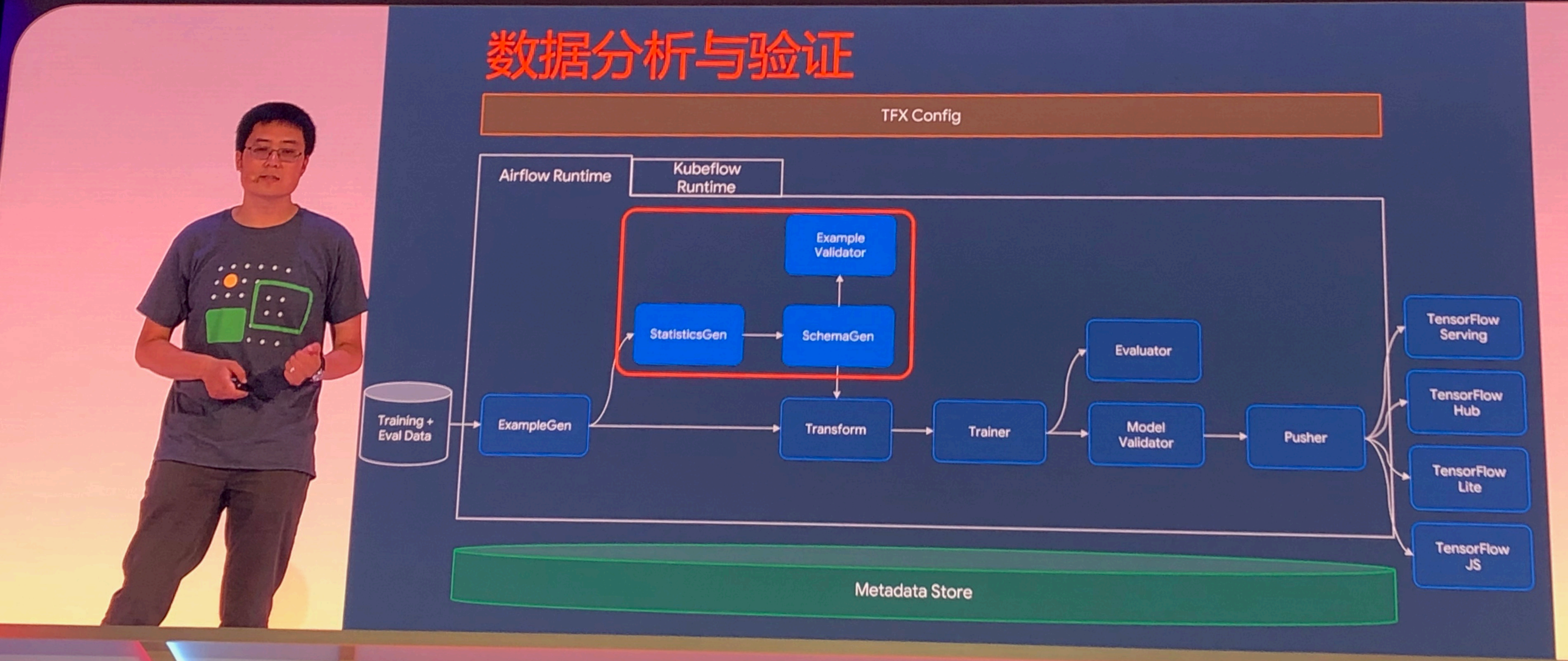
数据验证的重要性

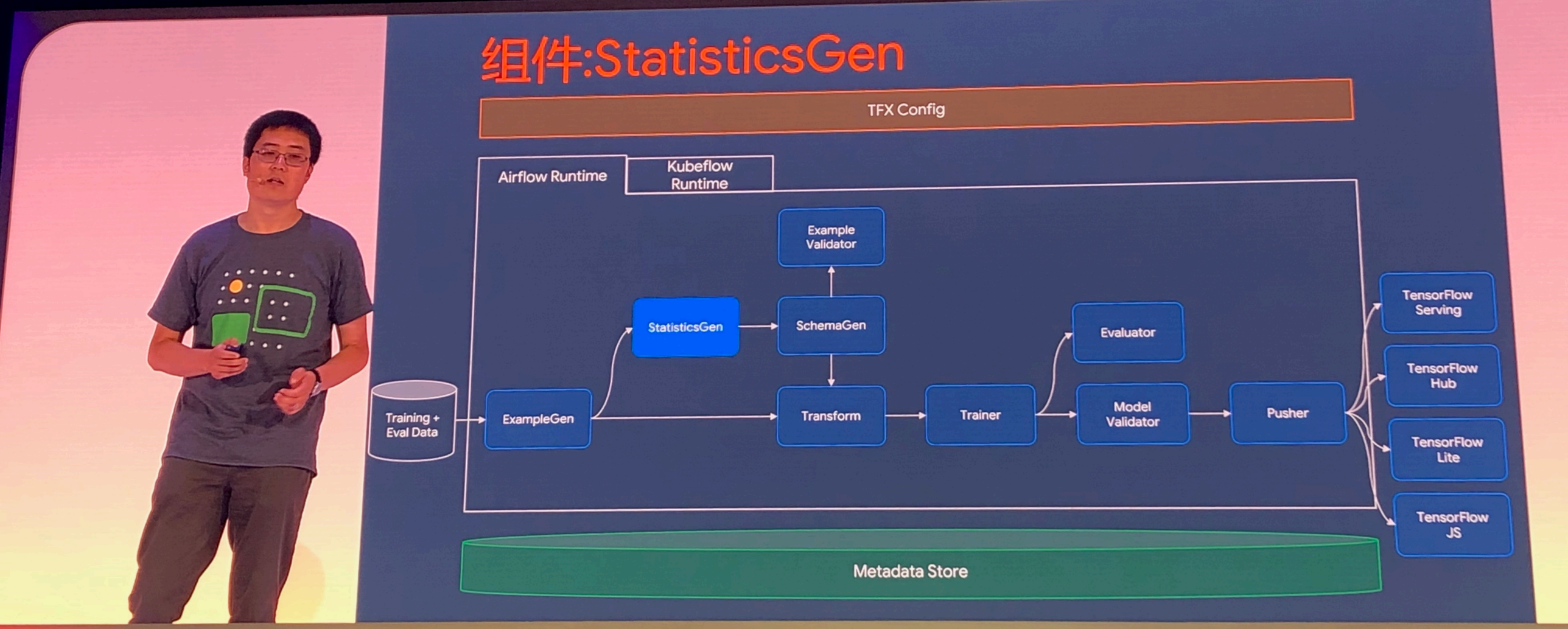


数据验证的重要性



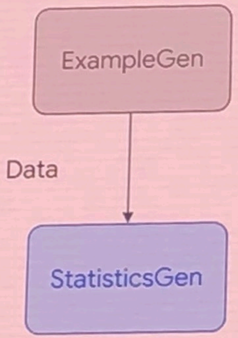
数据分析与验证


TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Training + Eval Data
ExampleGen
Transform
Trainer
Evaluator
Model Validator
Pusher
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Metadata Store


组件:StatisticsGen
TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Evaluator
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Training + Eval Data
ExampleGen
Transform
Trainer
Model Validator
Pusher
Metadata Store

组件: StatisticsGen

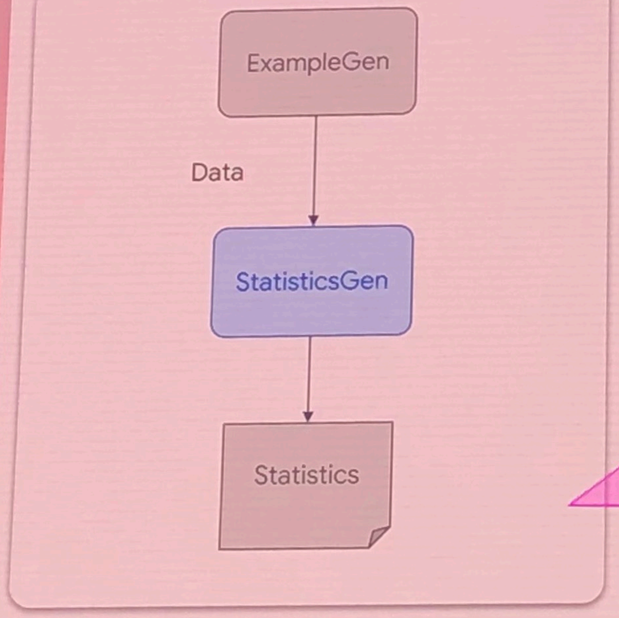
输入输出



- 训练数据
- 评价数据

组件: StatisticsGen

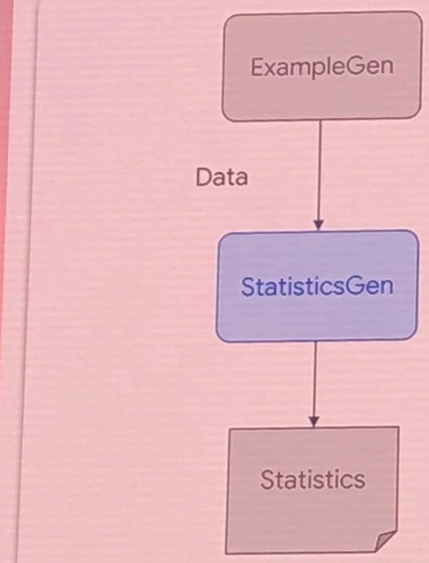
输入输出



- 数据分布
- 可视化

组件: StatisticsGen

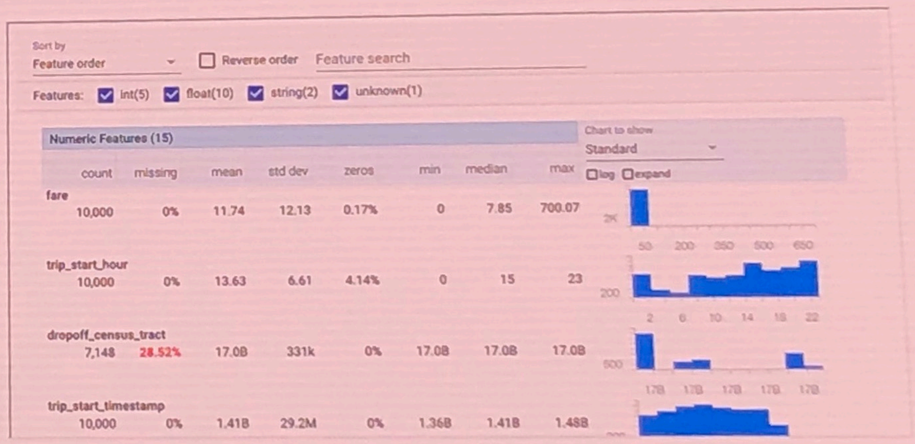
输入输出



配置

```
statistics_gen =
StatisticsGen(input_data=example_gen.outputs.examples)
```

可视化



凌晨的数据量偏少？

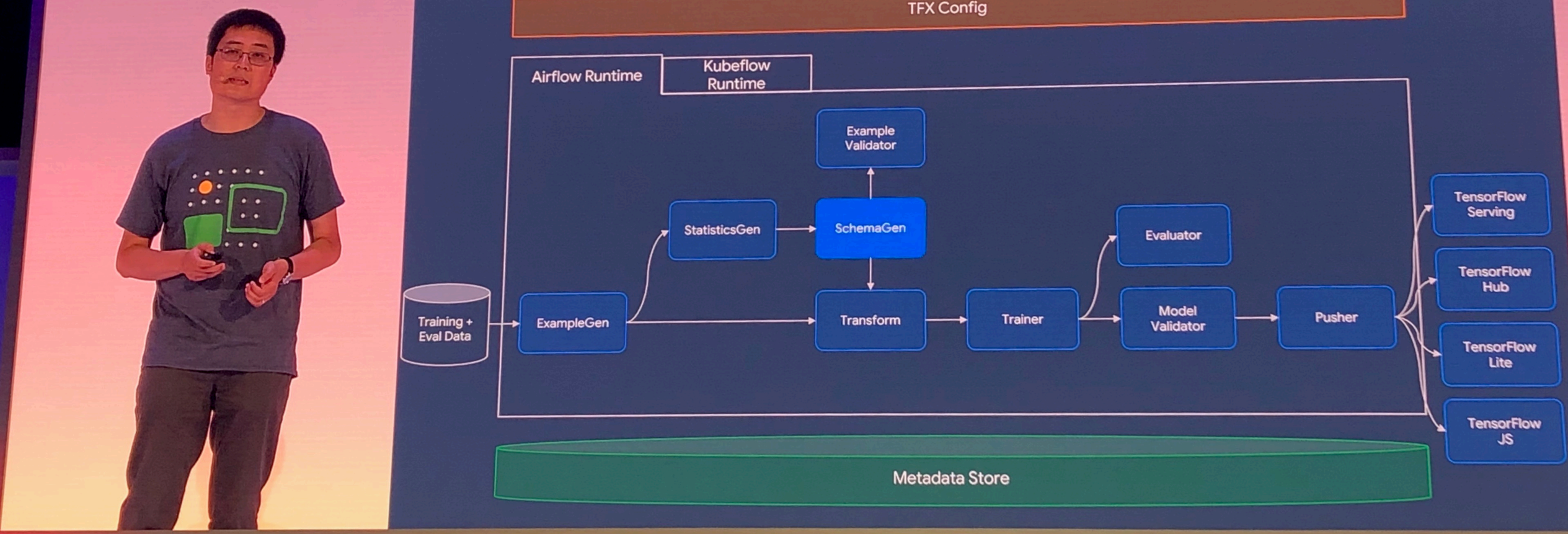
Sort by: Feature order | Reverse order | Feature search (regex enabled)

Features: int(5) | float(5) | variable-length floats(5) | string(1) | fixed-length strings(1) | variable-length strings(1)

Numeric Features (15) — Chart to show: Standard (log, expand)

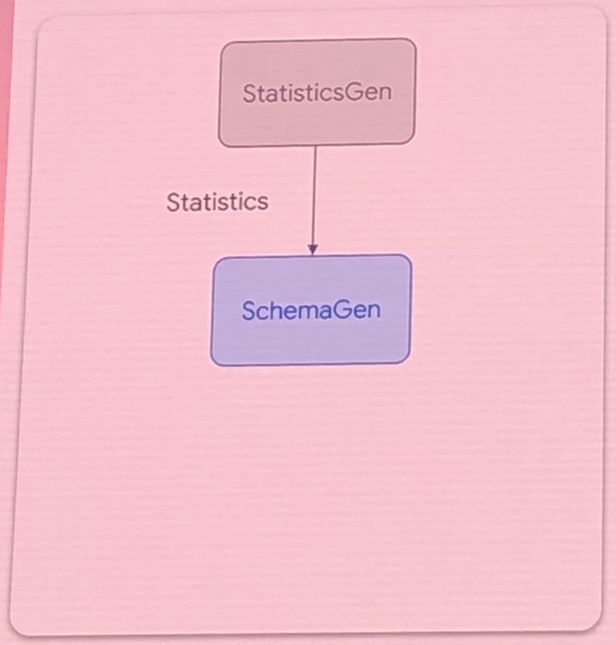
	count	missing	mean	std dev	zeros	min	median	max
fare	5,009	0%	11.97	14.12	0.12%	0	7.85	700.07
trip_start_hour	5,009	0%	13.51	6.73	4.13%	0	15	23
dropoff_census_tract	5,009	0%	17.0B	328k	0%	17.0B	17.0B	17.0B
trip_start_timestamp	5,009	0%	1.41B	29.0M	0%	1.36B	1.41B	1.48B
pickup_longitude	5,009	0%	-87.66	0.07	0%	-87.91	-87.63	-87.57
trip_start_month	5,009	0%	6.6	3.4	0%	1	7	12

组件:SchemaGen



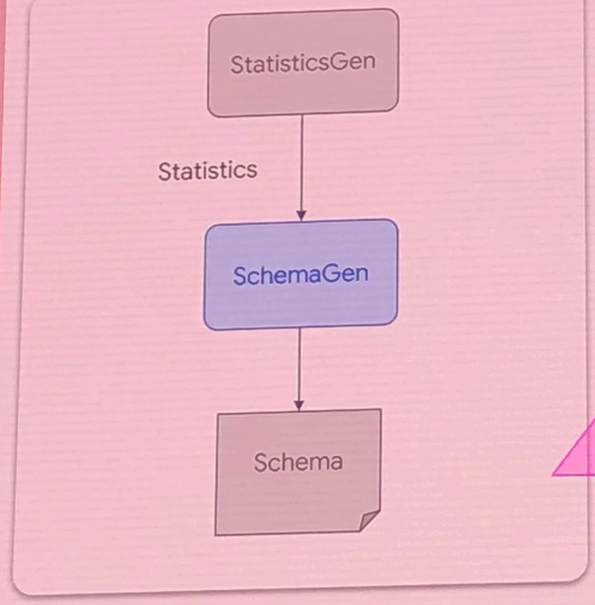
组件: SchemaGen

输入输出



组件: SchemaGen

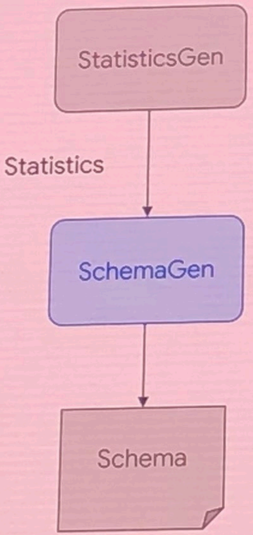
输入输出



- 数据的总体描述
  - 特征
  - 值域
  - 约束
  - 其它信息
- 结构化定义“正确”数据

组件: SchemaGen

输入输出



配置

```
infer_schema = SchemaGen(stats=statistics_gen.outputs.output)
```

可视化

Feature name	Type	Presence	Valency	Domain
'fare'	FLOAT	required	single	-
'trip_start_hour'	INT	required	single	-
'pickup_census_tract'	BYTES	optional		-
'dropoff_census_tract'	FLOAT	optional	single	-
'company'	STRING	optional	single	'company'
'trip_start_timestamp'	INT	required	single	-

付款方式有哪些合法取值?

Feature name				
'fare'	FLOAT	required	single	-
'trip_start_hour'	INT	required	single	-
'pickup_census_tract'	BYTES	optional		-
'dropoff_census_tract'	FLOAT	optional	single	-
'company'	STRING	optional	single	'company'
'trip_start_timestamp'	INT	required	single	-
'pickup_longitude'	FLOAT	required	single	-
'trip_start_month'	INT	required	single	-
'trip_miles'	FLOAT	required	single	-
'dropoff_longitude'	FLOAT	optional	single	-
'dropoff_community_area'	FLOAT	optional	single	-
'pickup_community_area'	INT	required	single	-
'payment_type'	STRING	required	single	'payment_type'
'trip_seconds'	FLOAT	optional	single	-
'trip_start_day'	INT	required	single	-
'tips'	FLOAT	required	single	-
'pickup_latitude'	FLOAT	required	single	-
'dropoff_latitude'	FLOAT	optional	single	-

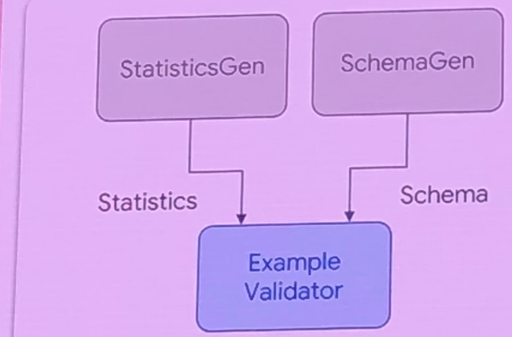
付款方式有哪些合法取值?

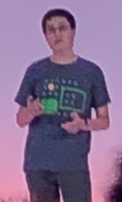
'trip_start_month'	INT	required	single	-
'trip_miles'	FLOAT	required	single	-
'dropoff_longitude'	FLOAT	optional	single	-
'dropoff_community_area'	FLOAT	optional	single	-
'pickup_community_area'	INT	required	single	-
'payment_type'	STRING	required	single	'payment_type'
'trip_seconds'	FLOAT	optional	single	-
'trip_start_day'	INT	required	single	-
'tips'	FLOAT	required	single	-
'pickup_latitude'	FLOAT	required	single	-
'dropoff_latitude'	FLOAT	optional	single	-

Domain	Values
'company'	'0118 - 42111 Godfrey S.Awir', '0694 - 59280 Chinesco Trans Inc', '1085 - 72312 N and W Cab Co', '2733 - 74600 Benny Jona', '2809 - 95474 C & D Cab Co Inc.', '3011 - 66308 JBL Cab Inc.', '3152 - 97284 Crystal Abernathy', '3201 - C&D Cab Co Inc', '3201 - CID Cab Co Inc', '3253 - 91138 Gaither Cab Co.', '3385 - 23210 Eman Cab', '3623 - 72222 Arrington Enterprises', '3897 - Ilie Malec', '4053 - Adwar H. Nikola', '4197 - 41842 Royal Star', '4615 - 83503 Tyrone Henderson', '4615 - Tyrone Henderson', '4623 - Jay Kim', '5006 - 39261 Salifu Bawa', '5006 - Salifu Bawa', '5074 - 54002 Ahzmi Inc', '5074 - Ahzmi Inc', '5129 - 87128', '5129 - 98755 Mengisti Taxi', '5129 - Mengisti Taxi', '5724 - KYVI Cab Inc', '585 - Valley Cab Co', '5864 - 73614 Thomas Owusu', '5864 - Thomas Owusu', '5874 - 73628 Sergey Cab Corp.', '5997 - 65283 AW Services Inc.', '5997 - AW Services Inc.', '6488 - 83287 Zuha Taxi', '6743 - Luhak Corp', 'Blue Ribbon Taxi Association Inc.', 'C & D Cab Co Inc', 'Chicago Elite Cab Corp.', 'Chicago Elite Cab Corp. (Chicago Carriag', 'Chicago Medallion Leasing INC', 'Chicago Medallion Management', 'Choice Taxi Association', 'Dispatch Taxi Affiliation', 'KOAM Taxi Association', 'Northwest Management LLC', 'Taxi Affiliation Services', 'Top Cab Affiliation'
'payment_type'	'Cash', 'Credit Card', 'Dispute', 'No Charge', 'Pcard', 'Unknown'

组件: ExampleValidator

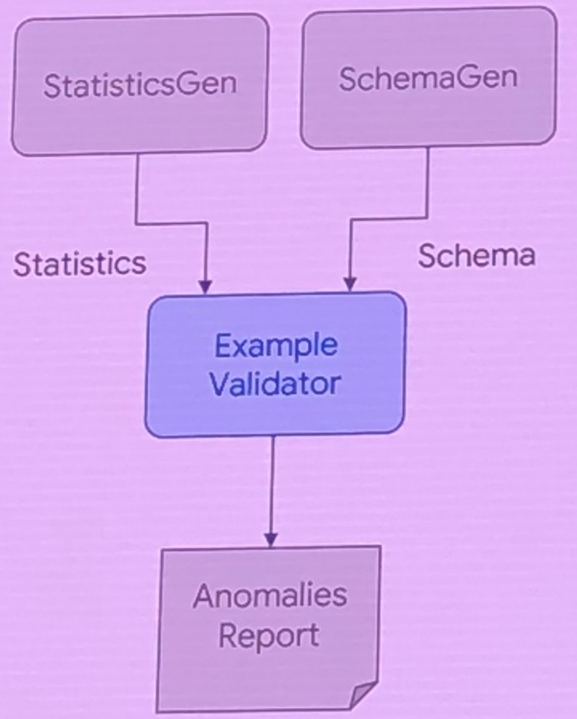
输入输出





组件: ExampleValidator

输入输出

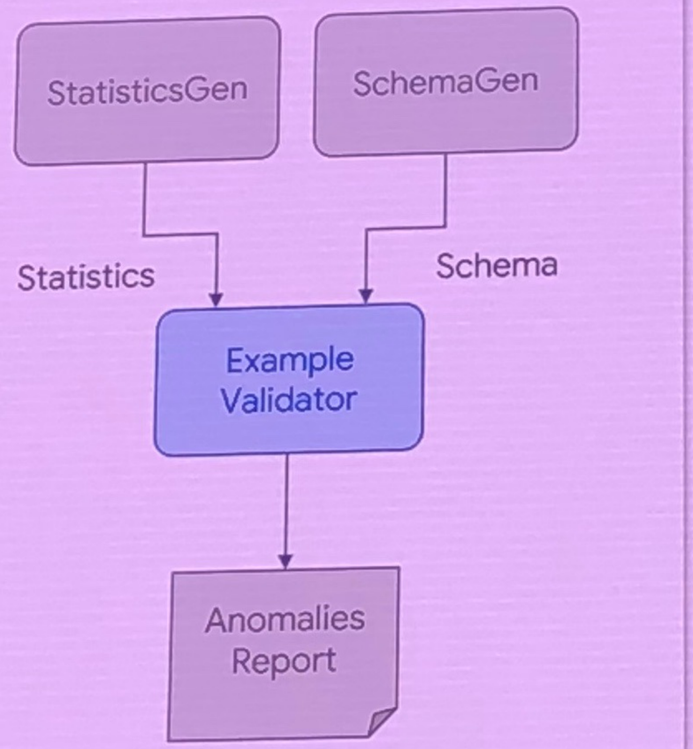


数据异常

- 缺少特征
- 未定义的特征值
- 训练/预测倾斜
- 未预料的数据分布变化
- ...

组件: ExampleValidator

输入输出



配置

```
validate_stats = ExampleValidator(
    stats=statistics_gen.outputs.output,
    schema=infer_schema.outputs.output)
```

可视化

Feature name	Anomaly short description	Anomaly long description
'payment_type'	Unexpected string values	Examples contain values missing from the schema: Prcard (<1%).
'company'	Unexpected string values	Examples contain values missing from the schema: 2092 - 61288 Sbeih company (<1%), 2192 - 73487 Zeymane Corp (<1%), 2192 - Zeymane Corp (<1%), 2823 - 73307 Seung Lee (<1%), 3094 - 24059 G.L.B. Cab Co (<1%), 3319 - CD Cab Co (<1%), 3385 - Eman Cab (<1%), 3897 - 57856 Ilie Malec (<1%), 4053 - 40193 Adwar H. Nikola (<1%), 4197 - Royal Star (<1%), 585 - 88805 Valley Cab Co (<1%), 5874 - Sergey Cab Corp. (<1%), 6057 - 24657 Richard Addo (<1%), 6574 - Babylon Express Inc. (<1%), 6742 - 83735 Tasha ride inc (<1%).

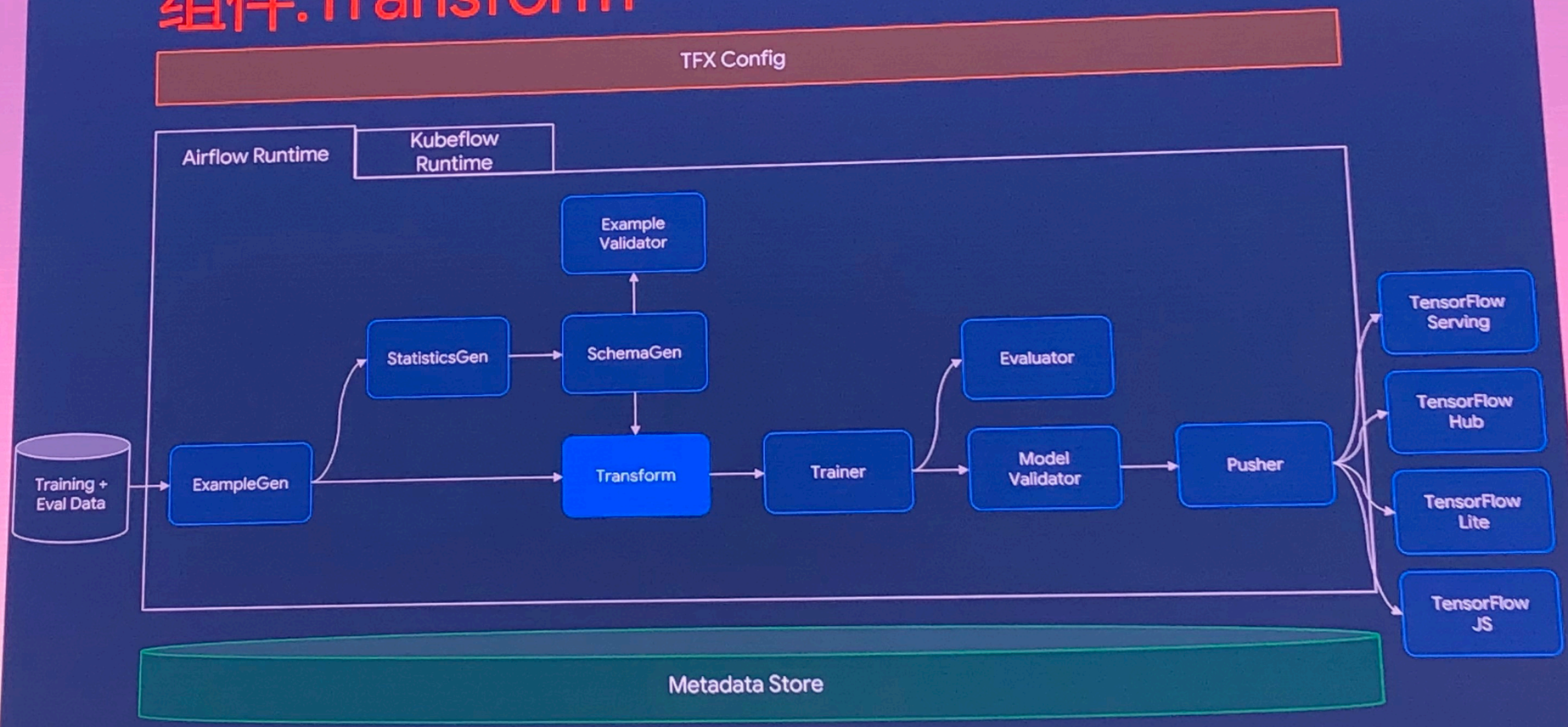
未定义的公司名：笔误还是新数据?

Feature name	Anomaly short description	Anomaly long description
'payment_type'	Unexpected string values	Examples contain values missing from the schema: Prcard (<1%).
'company'	Unexpected string values	Examples contain values missing from the schema: 2092 - 61288 Sbeih company (<1%), 2192 - 73487 Zeymane Corp (<1%), 2192 - Zeymane Corp (<1%), 2823 - 73307 Seung Lee (<1%), 3094 - 24059 G.L.B. Cab Co (<1%), 3319 - CD Cab Co (<1%), 3385 - Eman Cab (<1%), 3897 - 57856 Ilie Malec (<1%), 4053 - 40193 Adwar H. Nikola (<1%), 4197 - Royal Star (<1%), 585 - 88805 Valley Cab Co (<1%), 5874 - Sergey Cab Corp. (<1%), 6057 - 24657 Richard Addo (<1%), 6574 - Babylon Express Inc. (<1%), 6742 - 83735 Tasha ride inc (<1%).

TensorFlow

数据变换(Data Transformation)

组件:Transform



前情提要：TFX完整实例

芝加哥出租车数据集

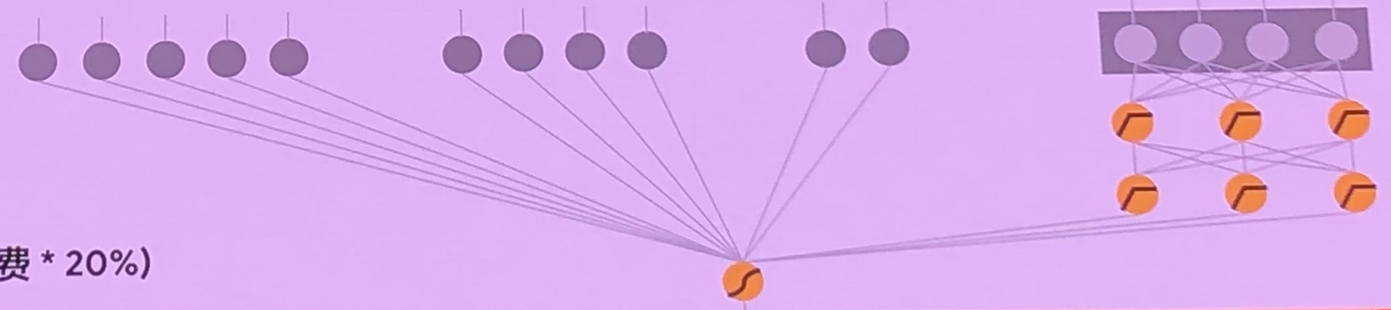
特征 (Features)

类别特征(Categorical)	离散化特征(Bucket)	字典化特征(Vocab)	数值特征(Dense Float)
trip_start_hour	pickup_latitude	payment_type	trip_miles
trip_start_day	pickup_longitude	company	fare
trip_start_month	dropoff_latitude		trip_seconds
pickup/dropoff_census_tract	dropoff_longitude		
pickup/dropoff_community_area			

变换 (Transforms)

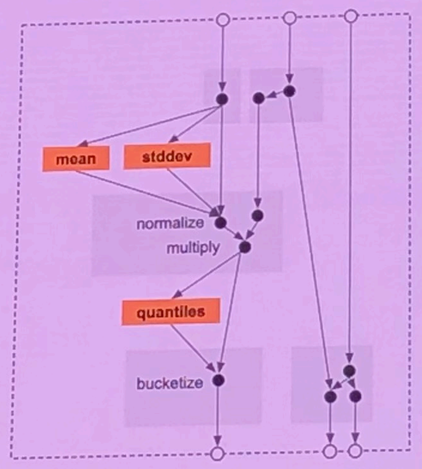
bucketize　string_to_int　scale_to_z_score

模型 (Wide+Deep)

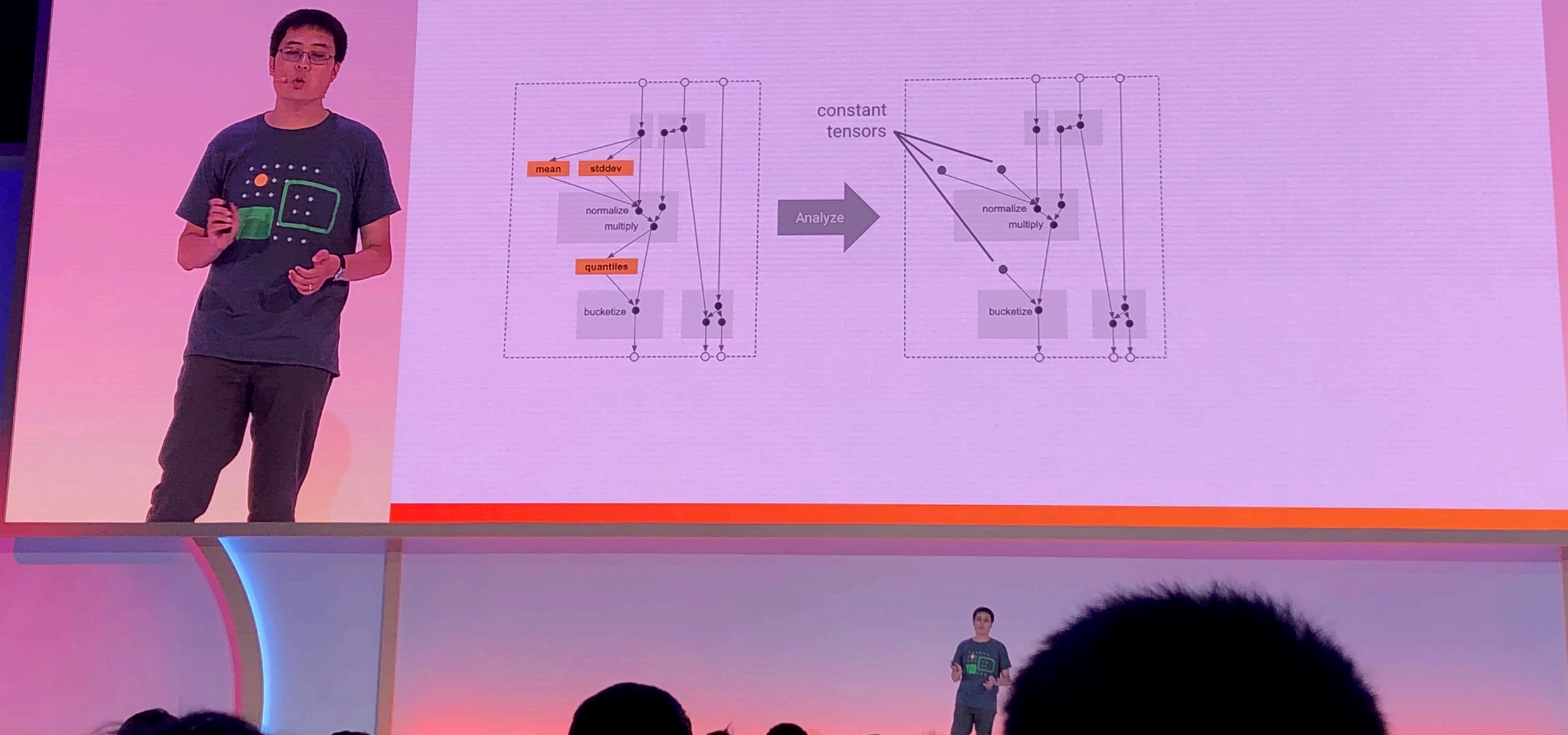


标签(Label) = 小费 > (车费 * 20%)

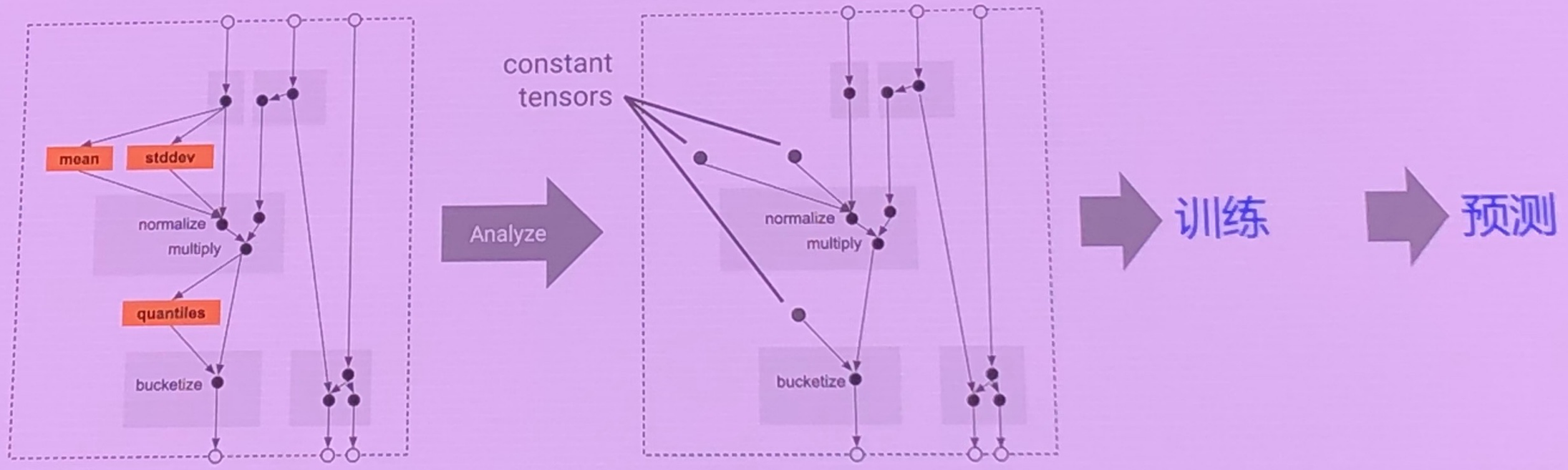
tf.Transform:特征变换



tf.Transform:特征变换



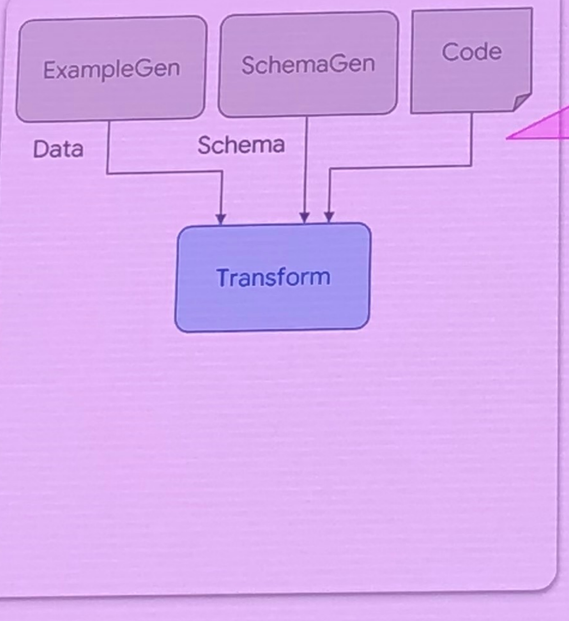
tf.Transform:特征变换



避免训练/预测倾斜(Train/Eval Skew)

组件: Transform

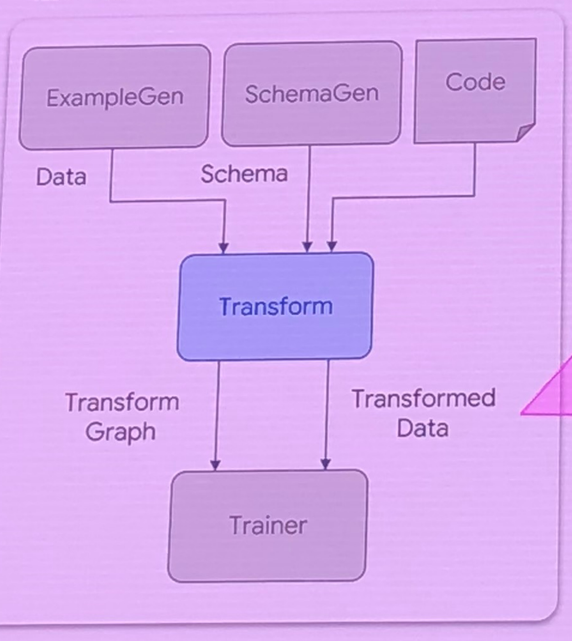
输入输出



- 用户自定义的变换代码
- 数据Schema

组件: Transform

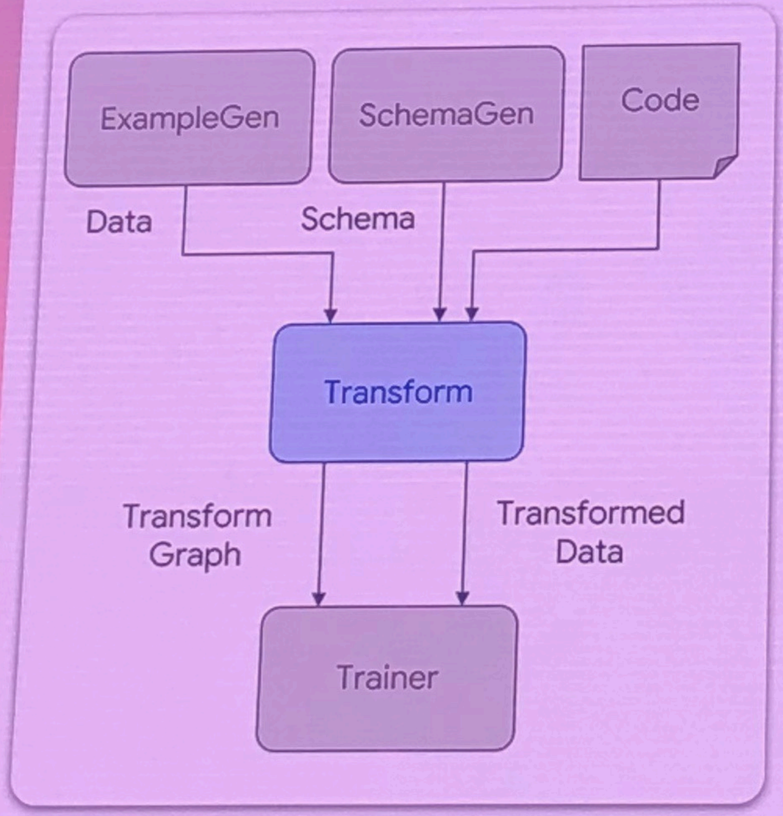
输入输出



- **变换图**
  - 可用于训练
  - 可内置于预测器
- **变换后数据**
  - 可选的训练器优化

组件: Transform

输入输出



配置

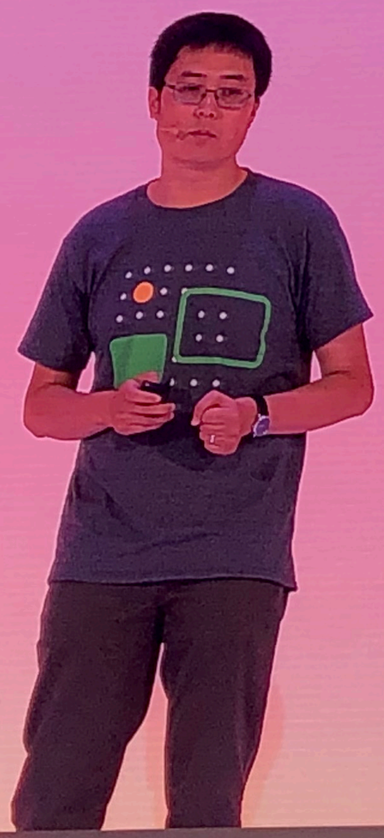
```
transform = Transform(
    input_data=example_gen.outputs.examples,
    schema=infer_schema.outputs.output,
    module_file=taxi_module_file)
```

变换代码

```
def preprocessing_fn(inputs):
  for key in taxi.DENSE_FLOAT_FEATURE_KEYS:
    outputs[key] = transform.scale_to_z_score(inputs[key])
  ...
  return outputs
```

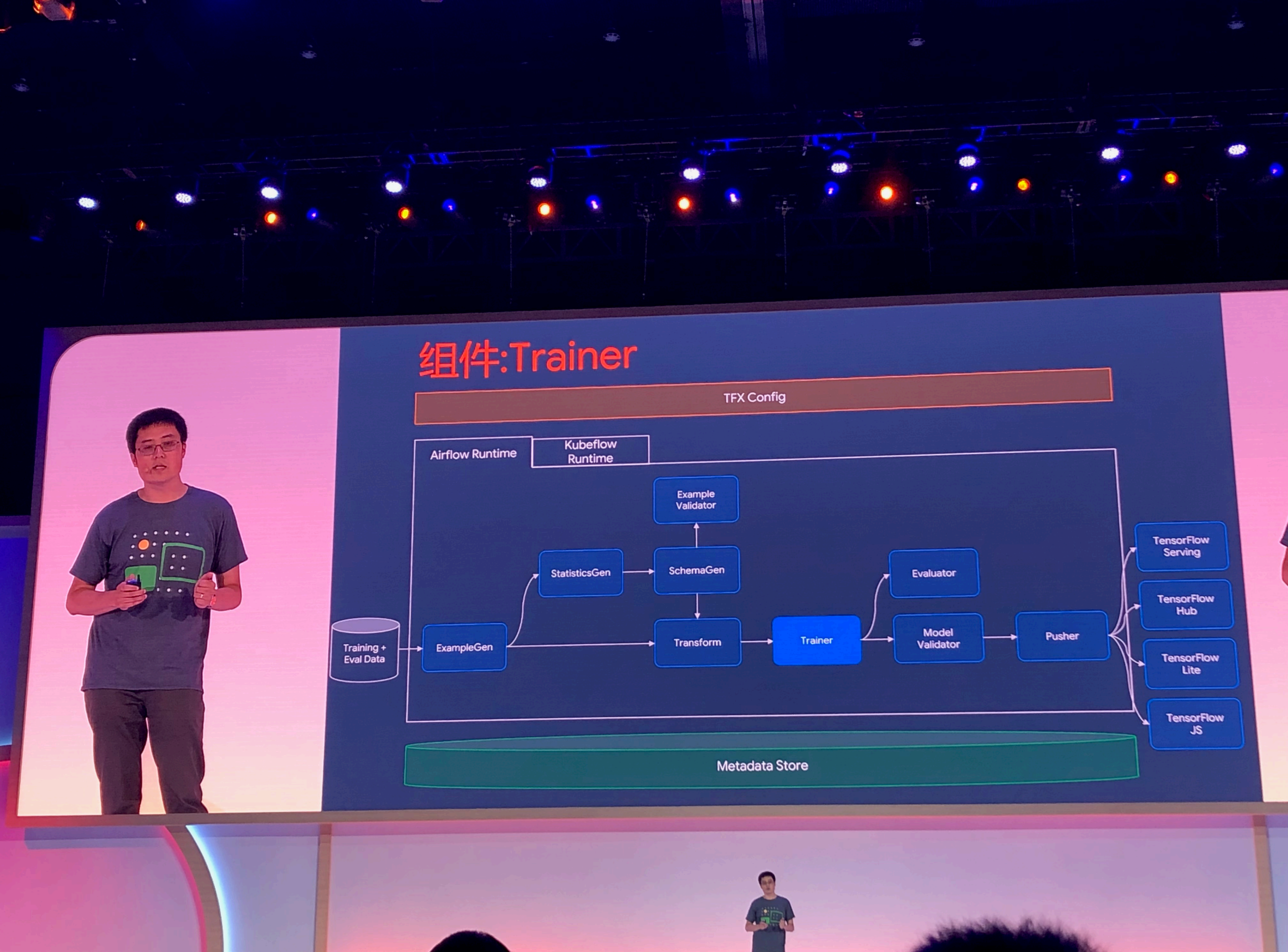
```
def preprocessing_fn(inputs):

  for key in taxi.DENSE_FLOAT_FEATURE_KEYS:
    outputs[key] = transform.scale_to_z_score(inputs[key])

  for key in taxi.VOCAB_FEATURE_KEYS:
    outputs[key] = transform.string_to_int(inputs[key],
        top_k=taxi.VOCAB_SIZE,
        num_oov_buckets=taxi.OOV_SIZE)

  for key in taxi.BUCKET_FEATURE_KEYS:
    outputs[key] = transform.bucketize(inputs[key],
        taxi.FEATURE_BUCKET_COUNT)

  return outputs
```

TensorFlow



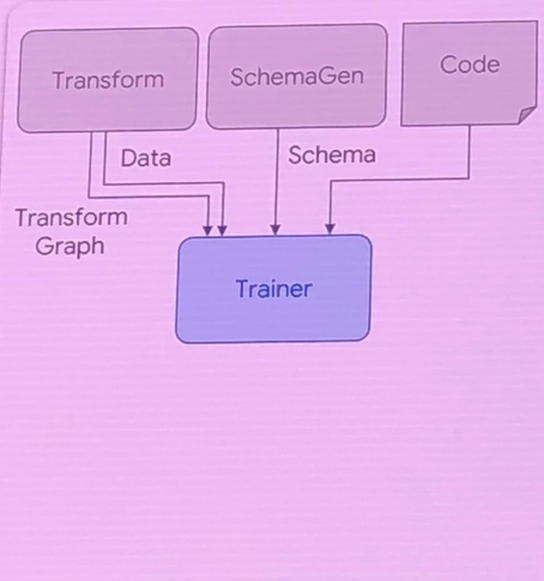
模型训练

组件:Trainer


TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Evaluator
Training + Eval Data
ExampleGen
Transform
Trainer
Model Validator
Pusher
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Metadata Store

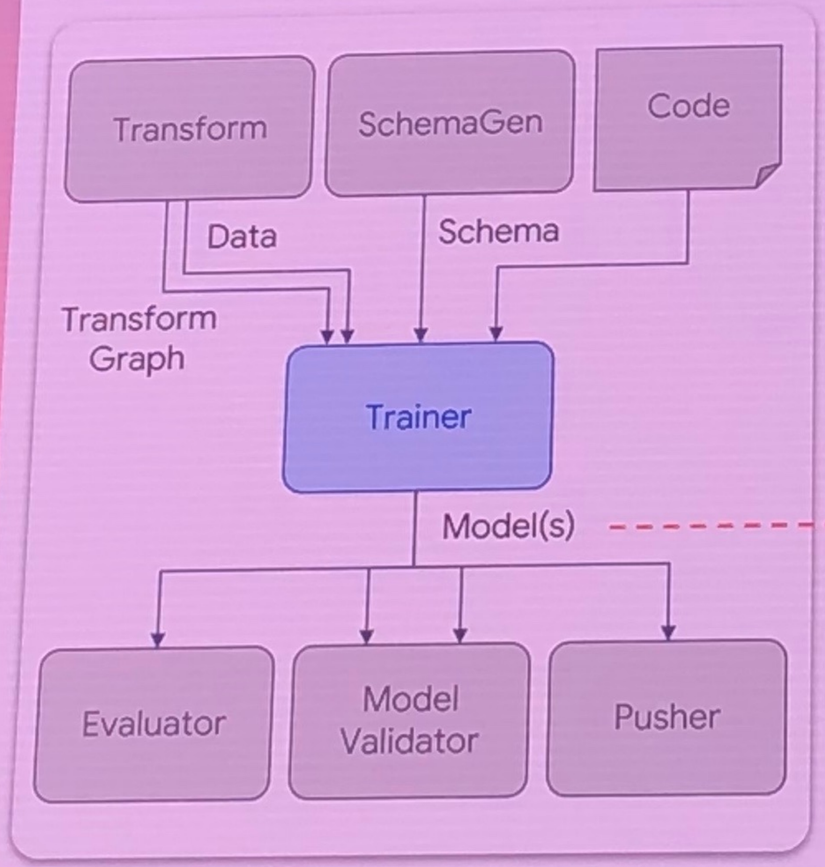
组件: Trainer

输入输出

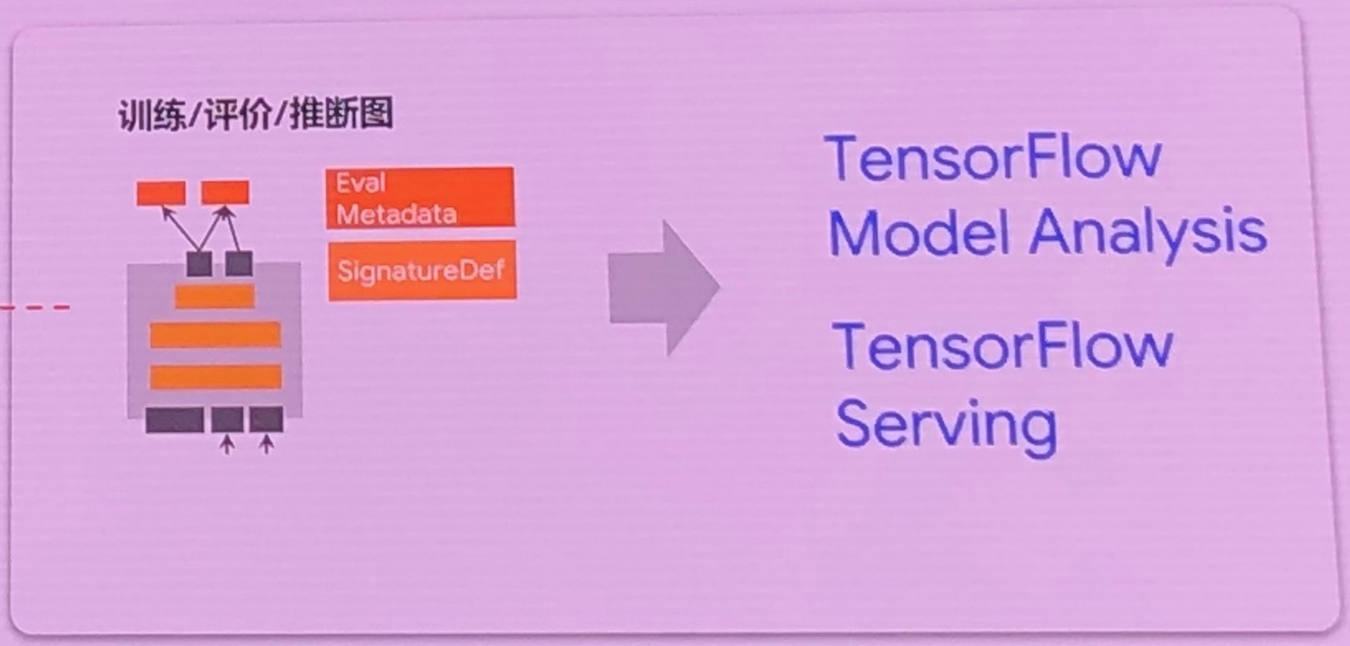


组件: Trainer

输入输出

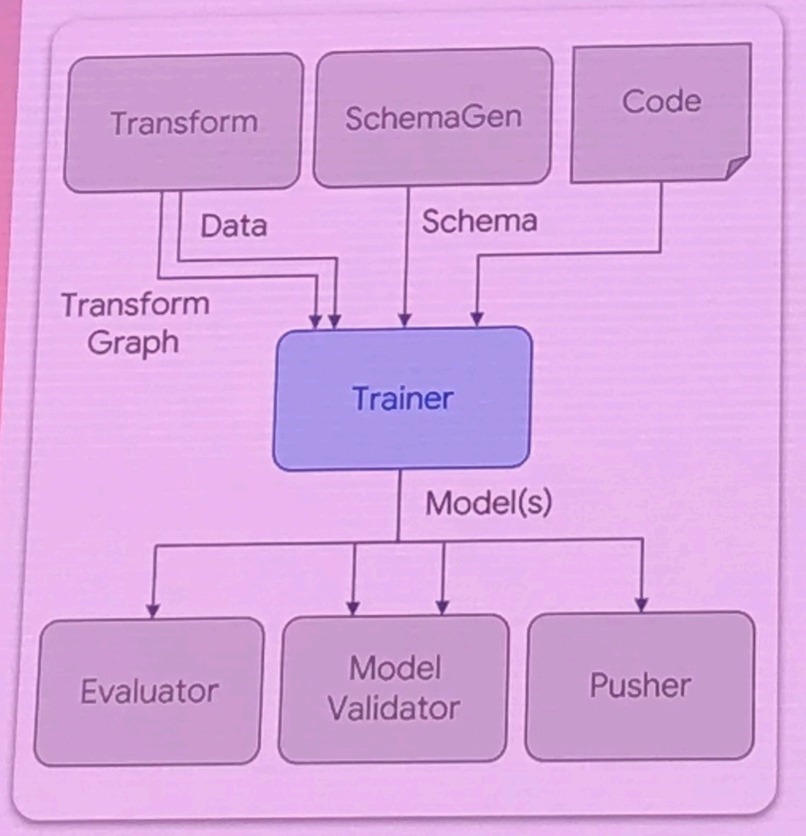


SavedModel：标准化的模型存储格式



组件: Trainer

输入输出



配置

```
trainer = Trainer(
    module_file=taxi_module_file,
    transformed_examples=transform.outputs.transformed_examples,
    schema=infer_schema.outputs.output,
    transform_output=transform.outputs.transform_output,
    train_steps=10000,
    eval_steps=5000,
    warm_starting=True)
```

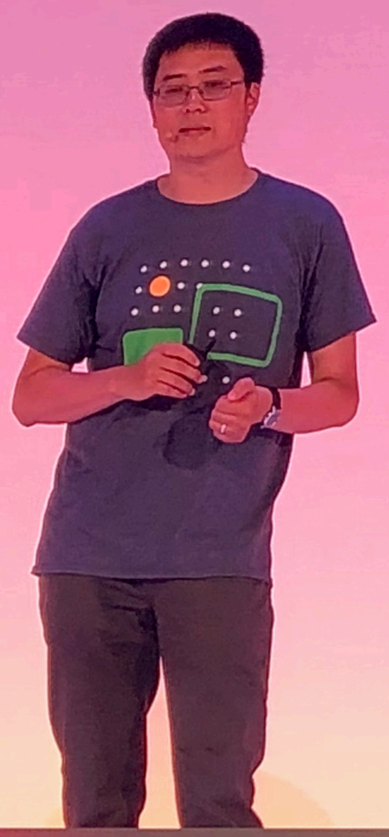
代码: 标准TensorFlow:)

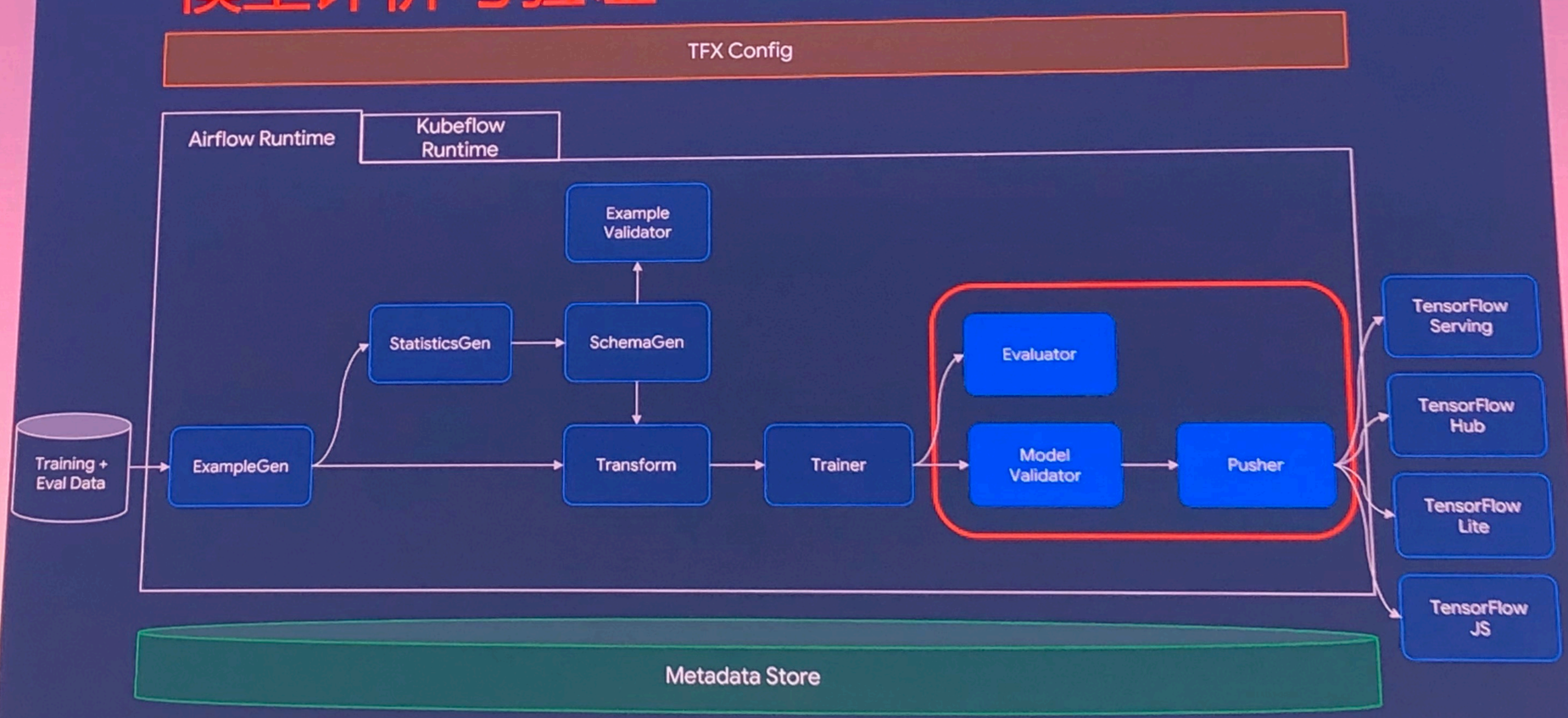
```python
def build_estimator(tf_transform_dir, config, hidden_units=None):

  # receive Schema and Transform metadata
  metadata_dir = os.path.join(tf_transform_dir,
                              transform_fn_io.TRANSFORMED_METADATA_DIR)
  transformed_metadata = metadata_io.read_metadata(metadata_dir)
  transformed_feature_spec = transformed_metadata.schema.as_feature_spec()

  transformed_feature_spec.pop(taxi.transformed_name(taxi.LABEL_KEY))

  real_valued_columns = [...]
 categorical_columns = [...]

 # Keras/TF 2.0 supported Soon!
 return tf.estimator.DNNLinearCombinedClassifier(
     config=config,
     linear_feature_columns=categorical_columns,
     dnn_feature_columns=real_valued_columns,
     dnn_hidden_units=hidden_units or [100, 70, 50, 25])
```

TensorFlow

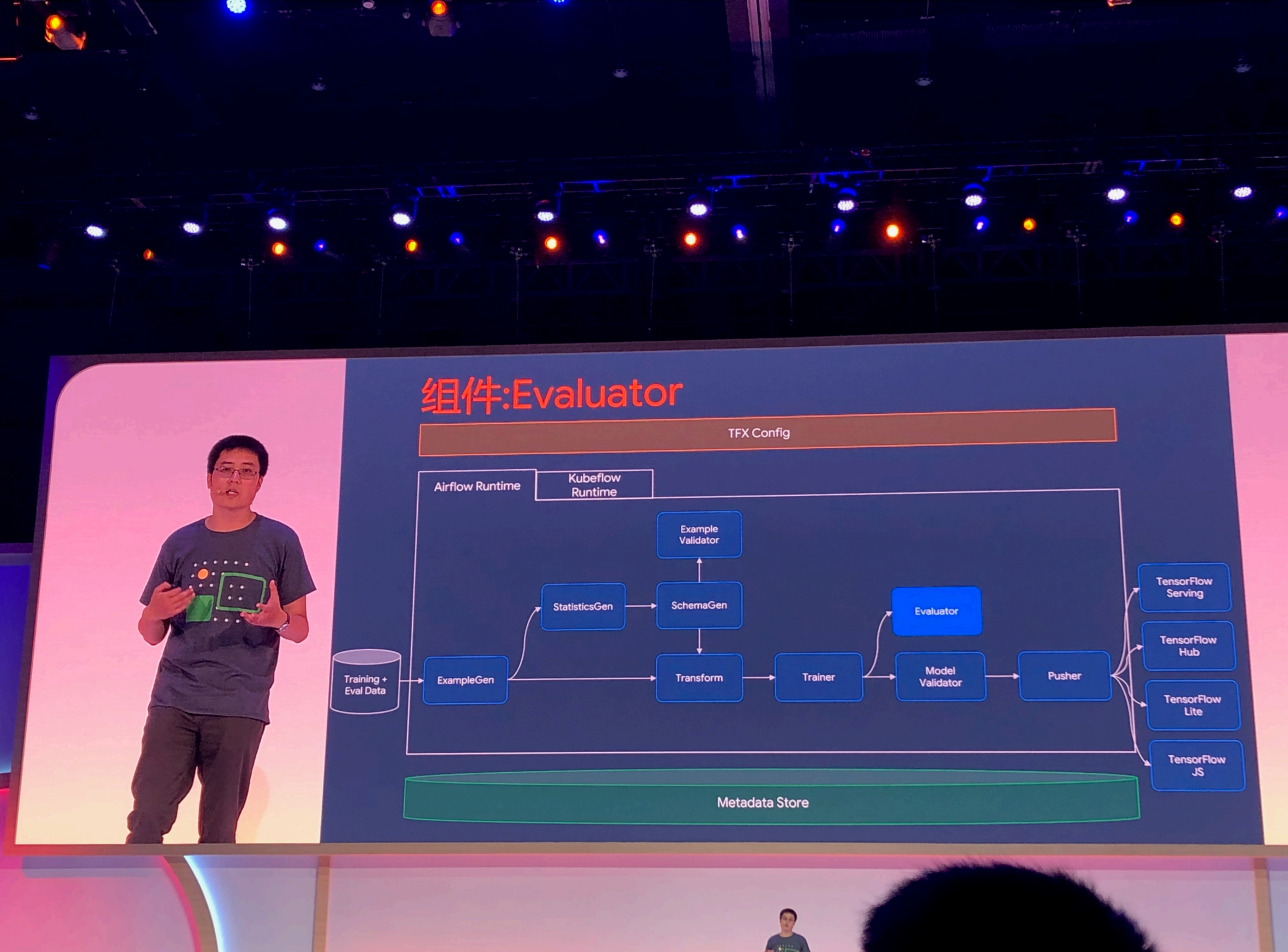
模型评价与验证

模型评价与验证



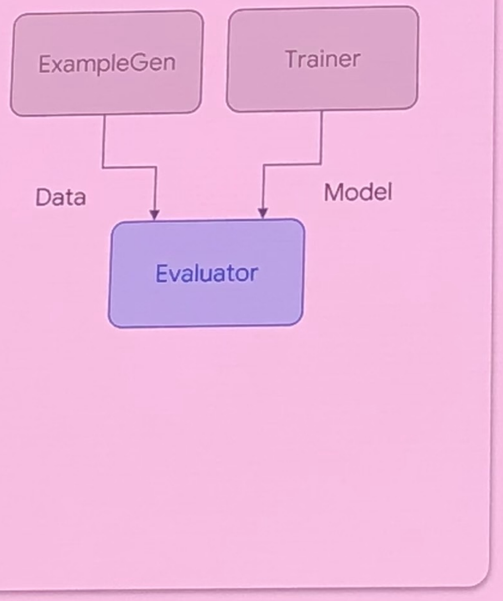

TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Training + Eval Data
ExampleGen
Transform
Trainer
Evaluator
Model Validator
Pusher
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Metadata Store

组件:Evaluator



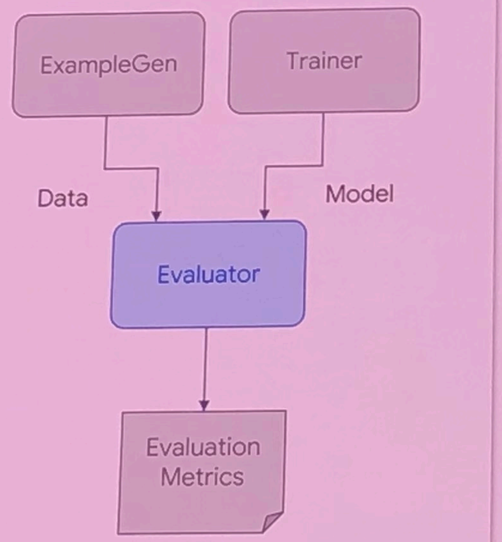
组件:Evaluator

输入输出



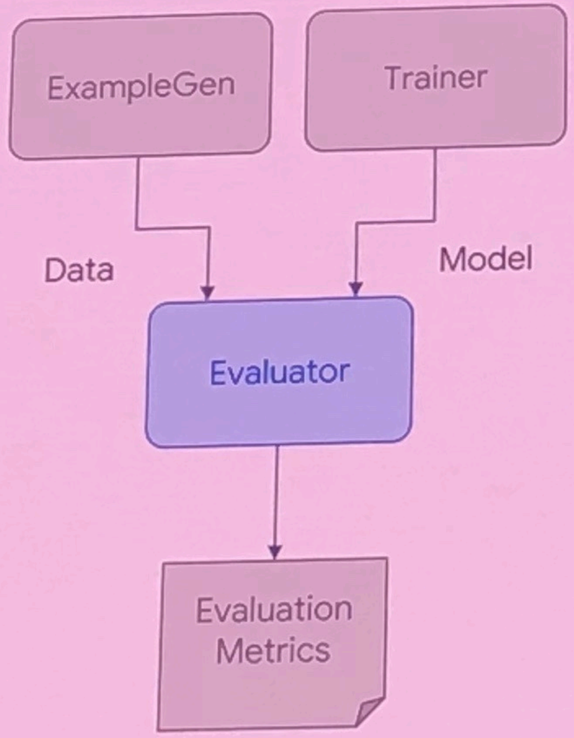
组件:Evaluator

输入输出



组件:Evaluator

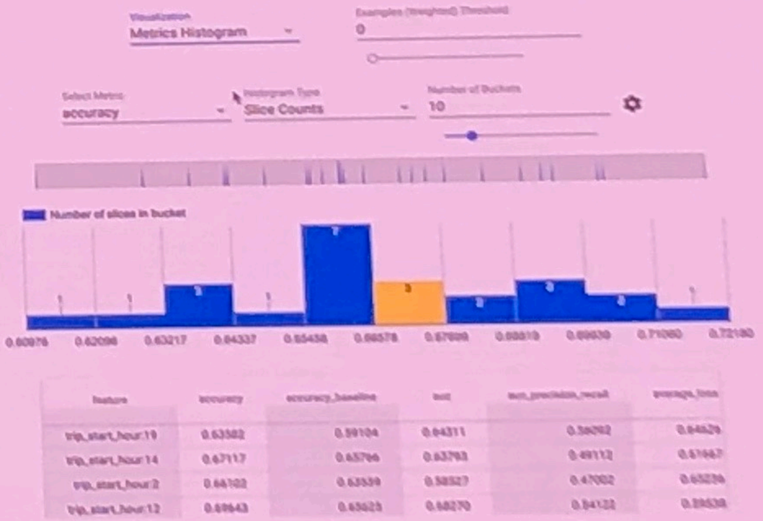
输入输出



配置

```
model_analyzer = Evaluator(
    examples=examples_gen.outputs.output,
    eval_spec=taxi_eval_spec,
    model_exports=trainer.outputs.output)
```

可视化



模型的总体性能?

Visualization: Slices Overview

Examples (Weighted) Threshold: 0

Show: post_export_metrics/example_count

Sort by: Slice

feature	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss	lab
Overall	0.94720	0.94220	0.92914	0.99480	0.14103	

模型在某个数据分片上的性能?

Visualization: Metrics Histogram

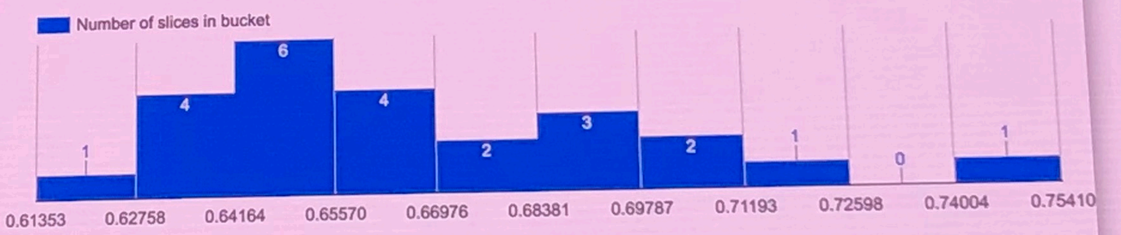
Examples (Weighted) Threshold

Select Metric: accuracy

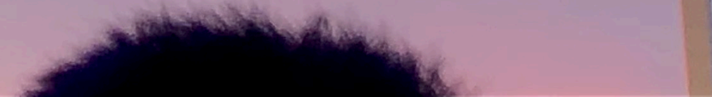
Histogram Type: Slice Counts

Number of Buckets: 10

Number of slices in bucket



feature	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss
trip_start_hour:19	0.65672	0.59104	0.66079	0.57315	0.64654
trip_start_hour:14	0.63964	0.65766	0.63072	0.46030	0.63655
trip_start_hour:2	0.64407	0.63559	0.55829	0.46379	0.67816
trip_start_hour:12	0.70536	0.65625	0.71230	0.57907	0.57703



模型在某个数据分片上的性能?

Slices Overview

Metrics Histogram

Examples (Weighted) Threshold

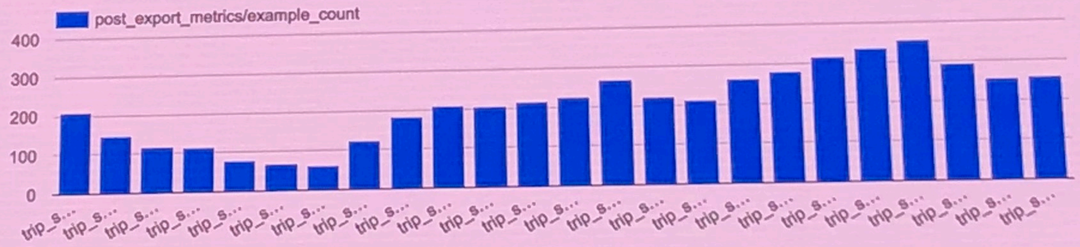
0

Show

post_export_metrics/example_count

Sort by

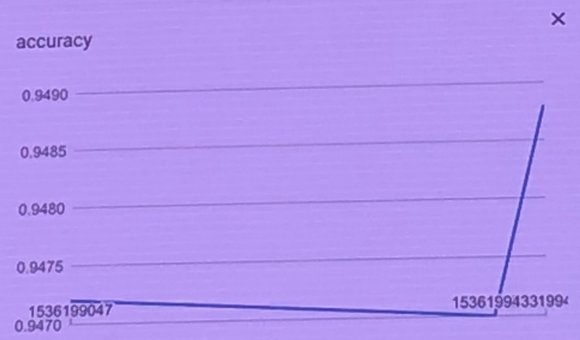
Slice

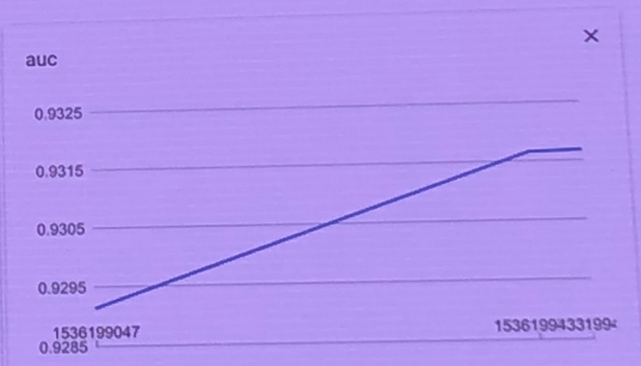


feature	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss
trip_start_hour:19	0.65672	0.59104	0.66079	0.57315	0.64654
trip_start_hour:14	0.63964	0.65766	0.63072	0.46030	0.63655
trip_start_hour:2	0.64407	0.63559	0.55829	0.46379	0.67816
trip_start_hour:12	0.70536	0.65625	0.71230	0.57907	0.57703
trip_start_hour:0	0.63768	0.66667	0.62093	0.42289	0.62715
trip_start_hour:23	0.66016	0.64844	0.58337	0.44173	0.65142

模型性能是否在随时间而改善?

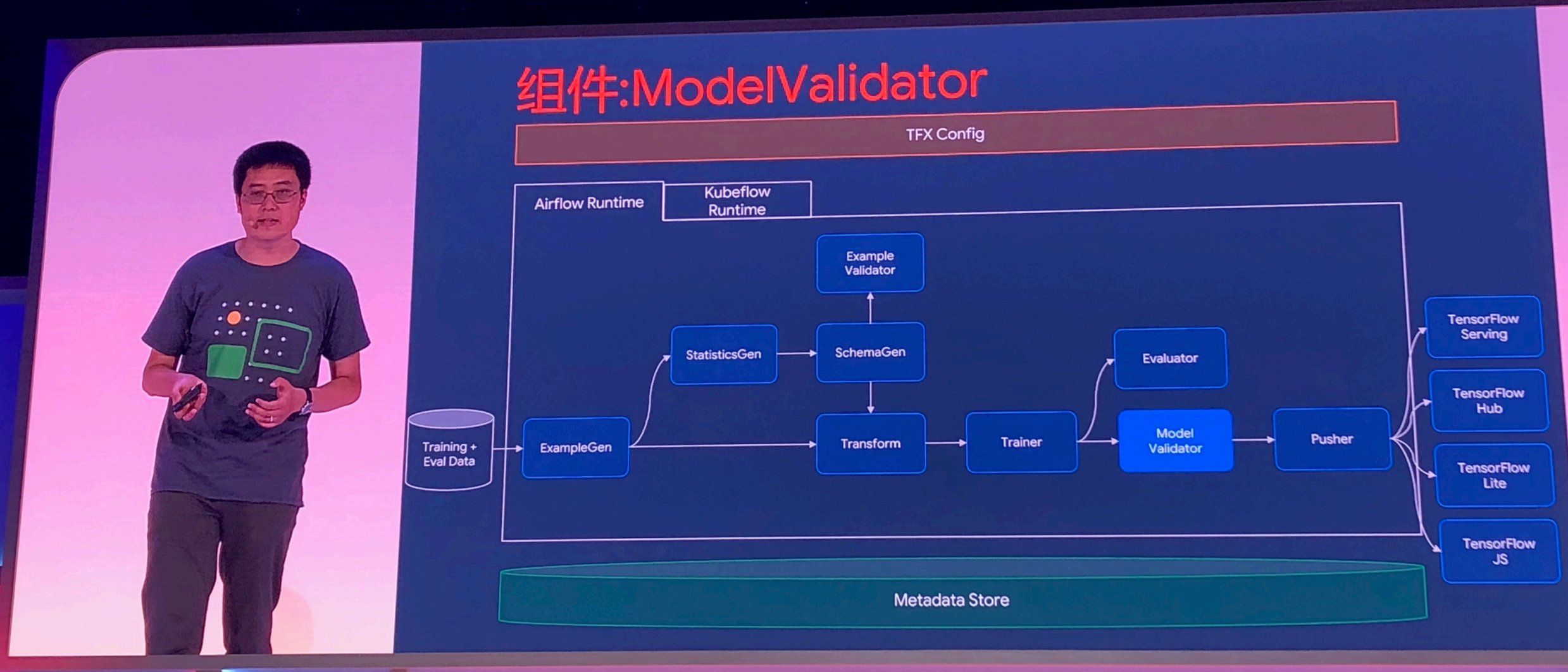
Add metric series





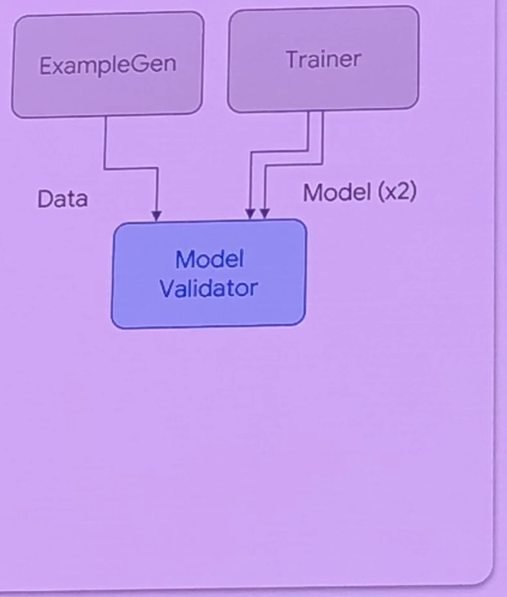
Model	Data	accuracy	accuracy_baseline	auc	auc_precision_recall	average_loss	label/mean	po
1536199479	data.csv	0.94880	0.94220	0.93168	0.98516	0.13980	0.94220	
1536199433	data.csv	0.94700	0.94220	0.93165	0.98170	0.13979	0.94220	
1536199047	data.csv	0.94720	0.94220	0.92914	0.99480	0.14103	0.94220	

组件:ModelValidator



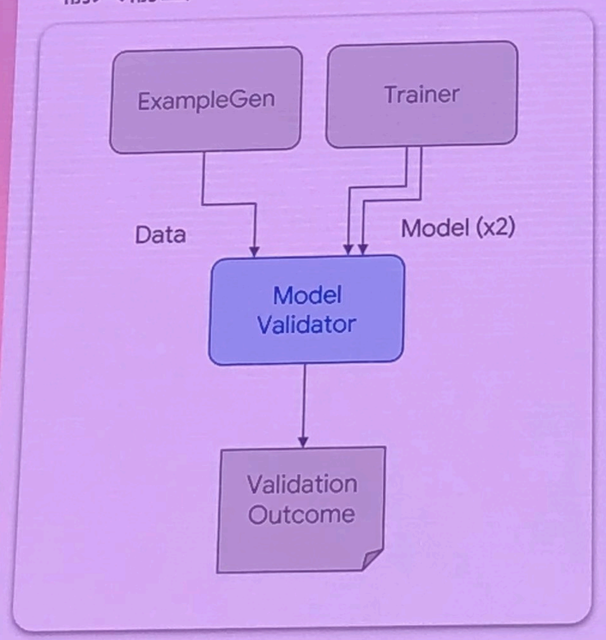
组件:ModelValidator

输入输出

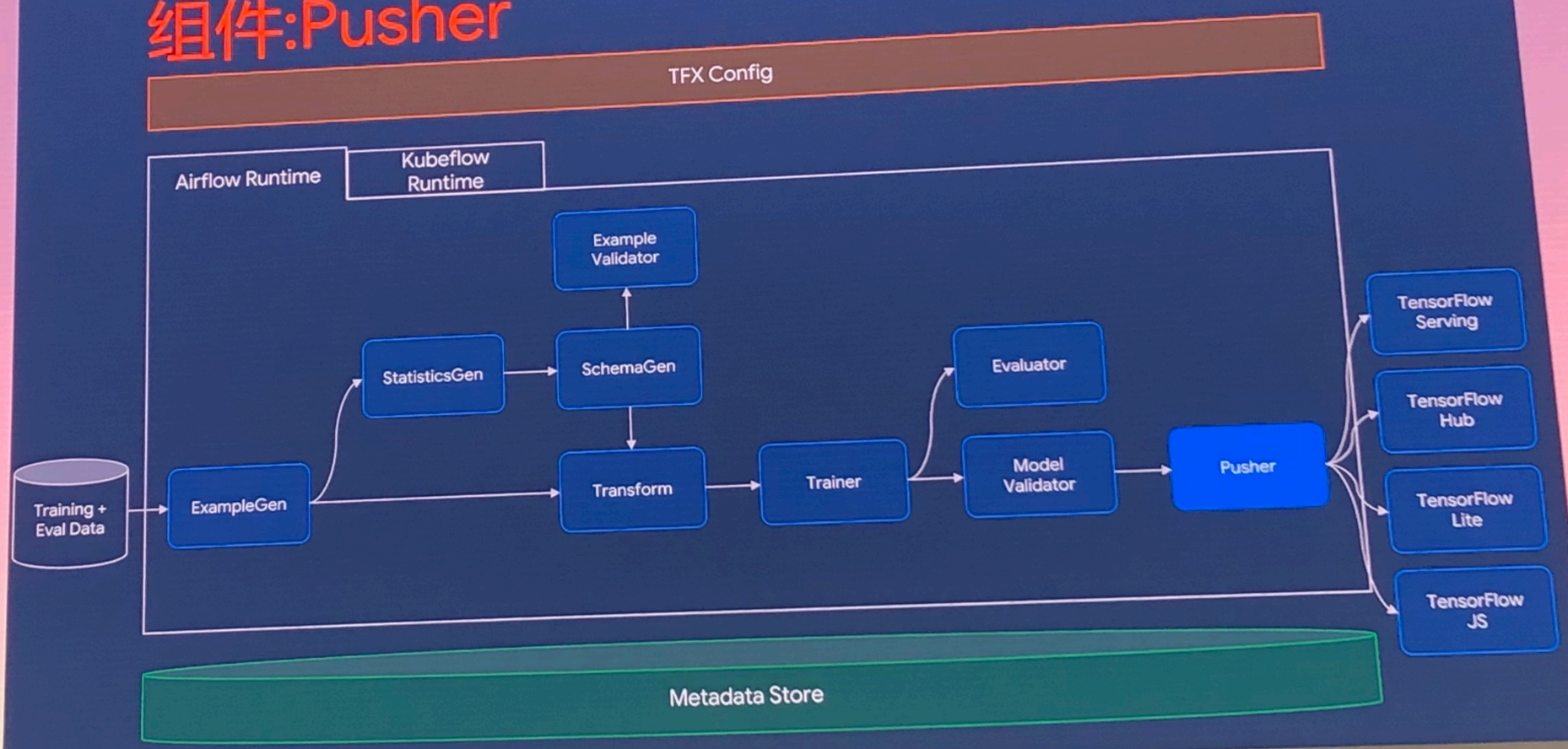


组件:ModelValidator

输入输出

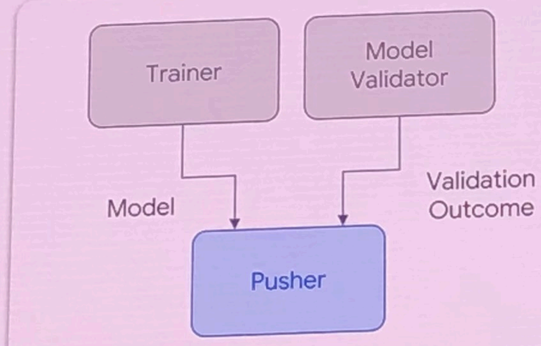


组件:Pusher


TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Evaluator
Training + Eval Data
ExampleGen
Transform
Trainer
Model Validator
Pusher
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Metadata Store

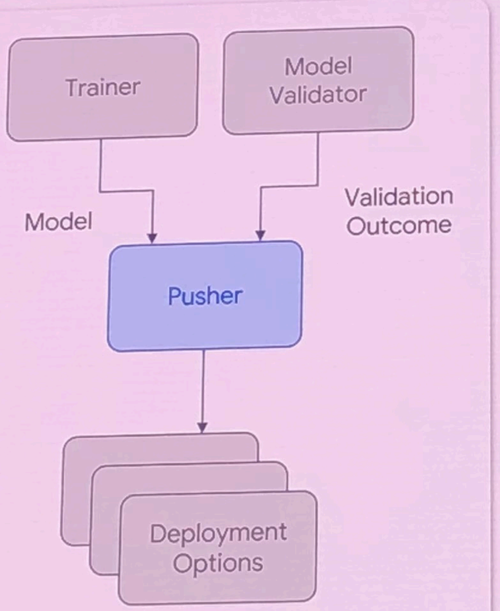
组件:Pusher

输入输出



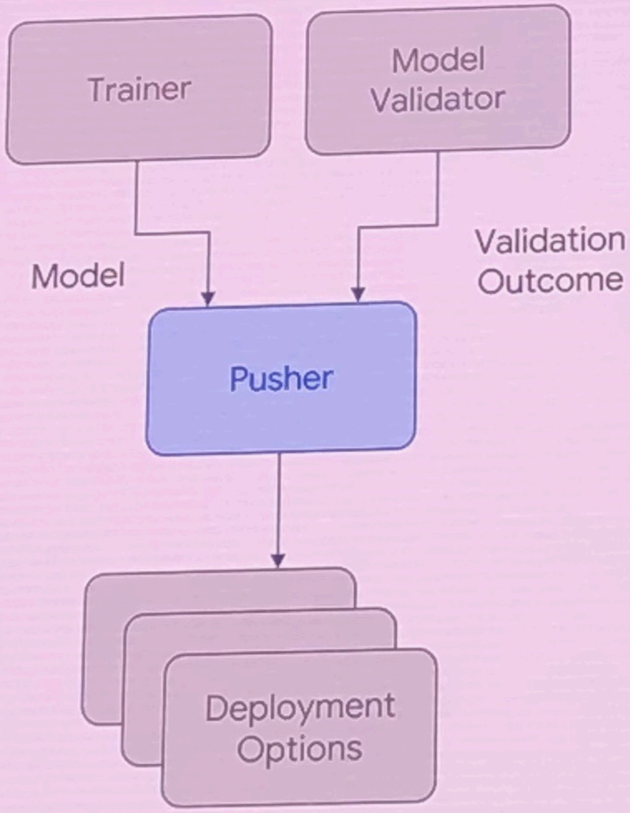
组件:Pusher

输入输出



组件:Pusher

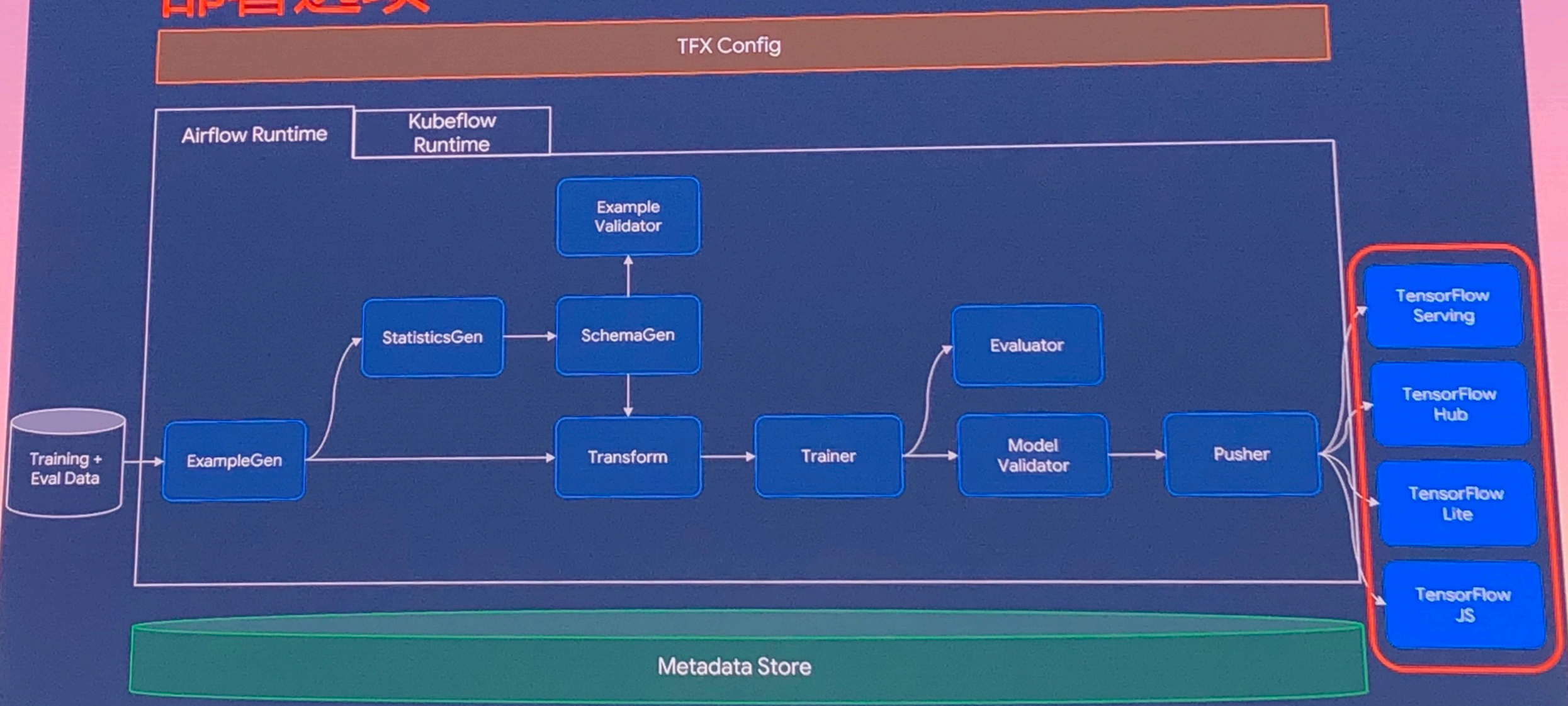
输入输出



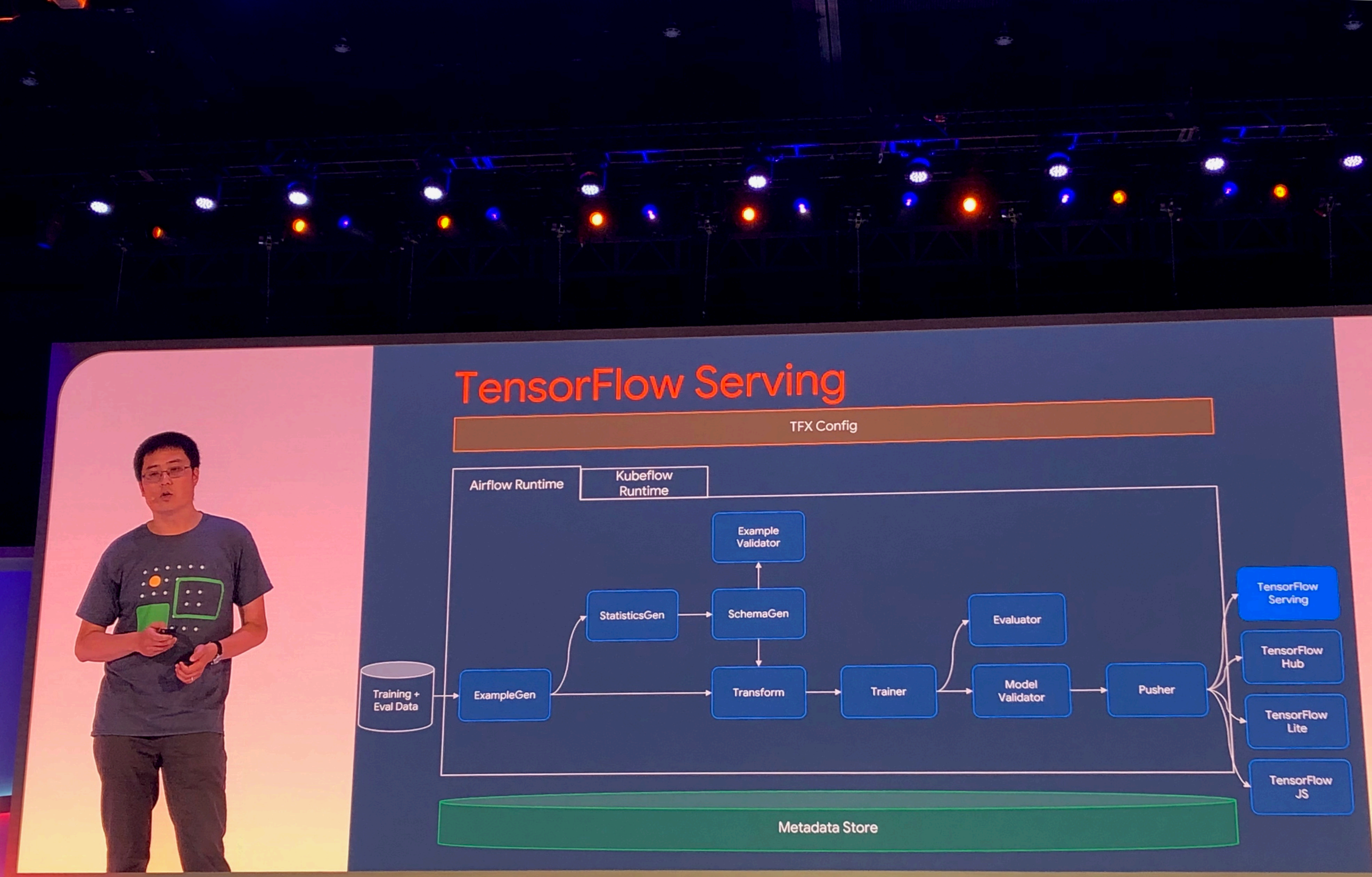
配置

```
pusher = Pusher(
    model_export=trainer.outputs.output,
    model_blessing=model_validator.outputs.blessing,
    serving_model_dir=serving_model_dir)
```

部署选项


TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Evaluator
Training + Eval Data
ExampleGen
Transform
Trainer
Model Validator
Pusher
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Metadata Store

TensorFlow Serving

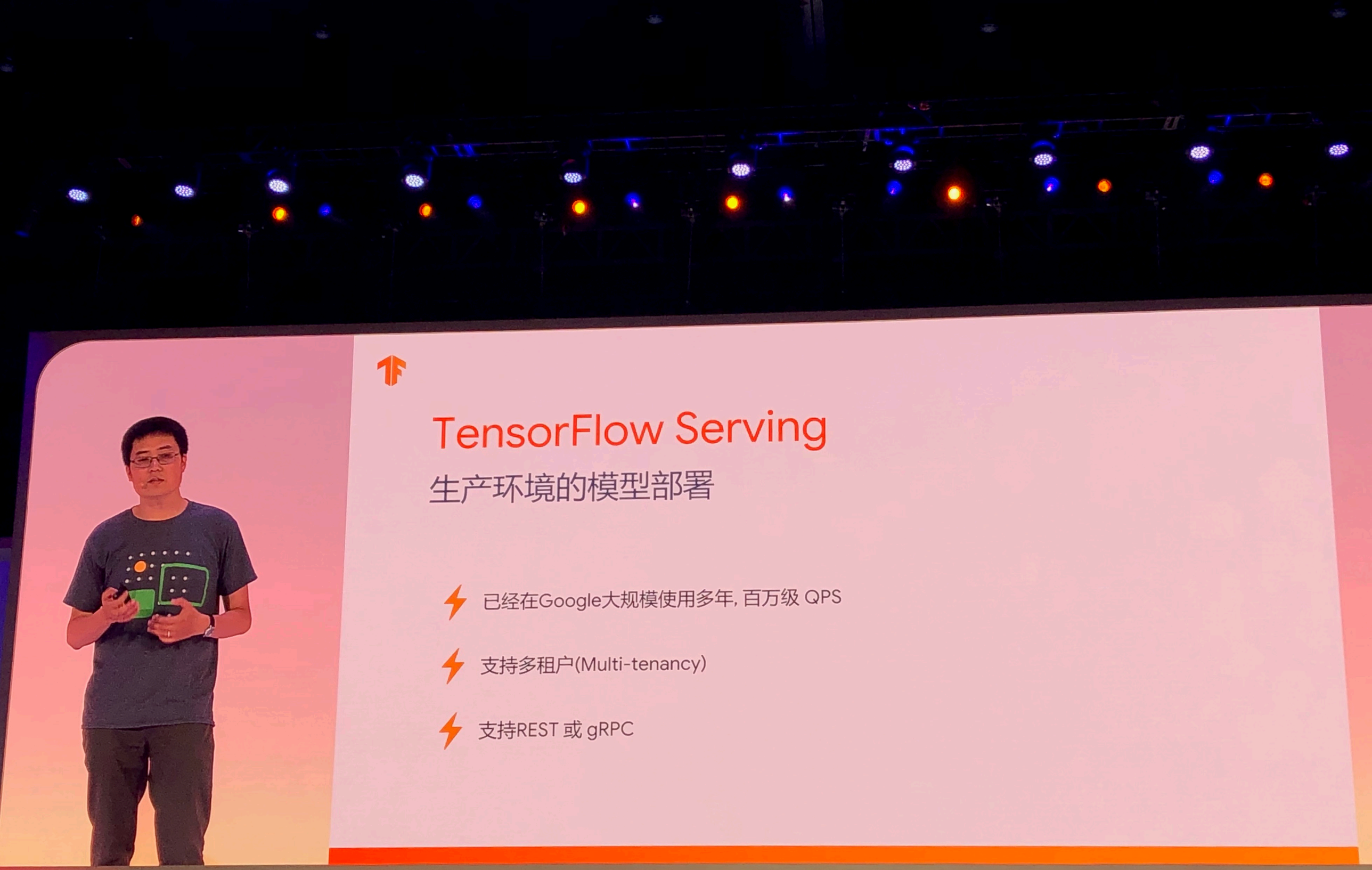


TensorFlow Serving
生产环境的模型部署
已经在Google大规模使用多年, 百万级 QPS

TensorFlow Serving

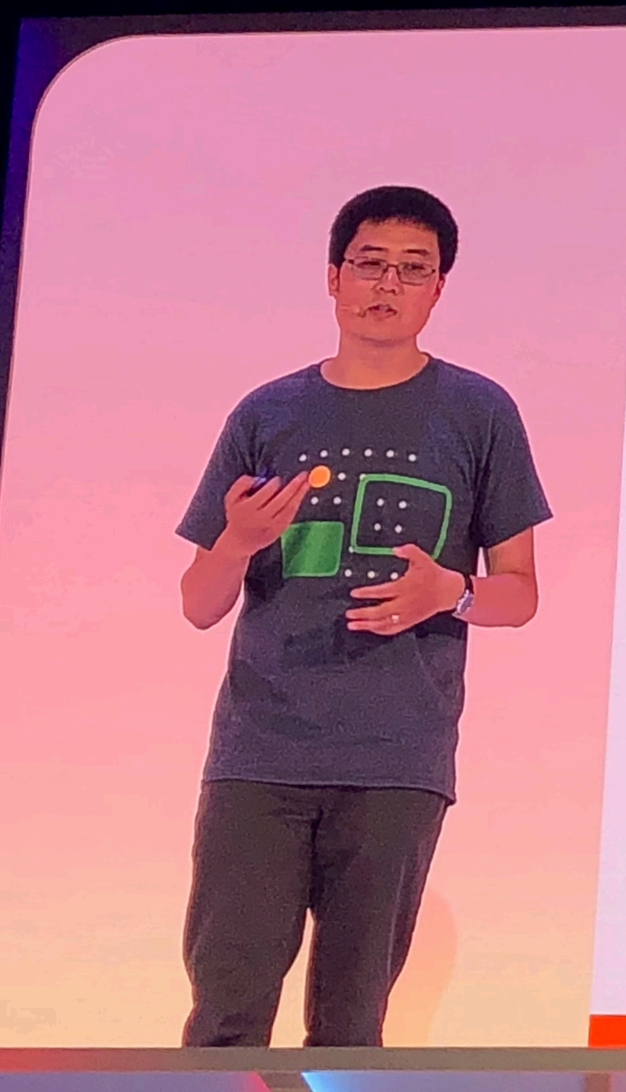
生产环境的模型部署

- 已经在Google大规模使用多年, 百万级 QPS
- 支持多租户(Multi-tenancy)


TensorFlow Serving
生产环境的模型部署
已经在Google大规模使用多年, 百万级 QPS
支持多租户(Multi-tenancy)
支持REST 或 gRPC

模型服务器：一键部署

```
$ tensorflow_model_server
    --port=8501
    --model_name=chicago_taxi
    --model_base_path='/path/to/savedmodel'
```



Docker容器：一键部署

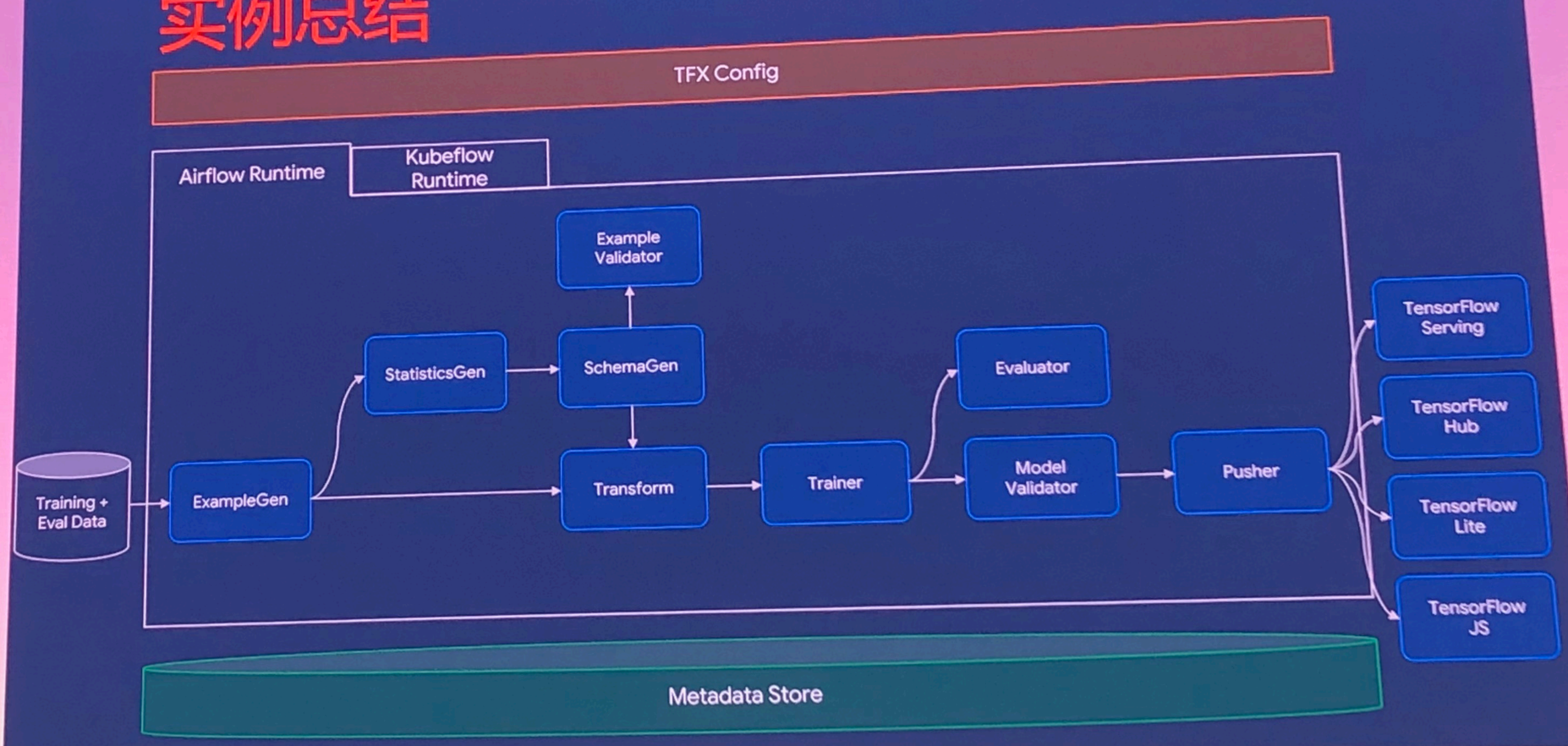
```
$ docker run -p 8501:8501 \
    -v '/path/to/savedmodel':/models/chicago_taxi
    -e MODEL_NAME=chicago_taxi -t tensorflow/serving
```

REST API预测

```
$ curl -d '{"instances": [60, 27.05, ...]}' \
  -X POST http://localhost:8501/v1/models/chicago_taxi:predict
```

```
{ "results": [1.0] }
```

实例总结


TFX Config
Airflow Runtime
Kubeflow Runtime
Example Validator
StatisticsGen
SchemaGen
Evaluator
Training + Eval Data
ExampleGen
Transform
Trainer
Model Validator
Pusher
TensorFlow Serving
TensorFlow Hub
TensorFlow Lite
TensorFlow JS
Metadata Store