

سامانه نرم افزاری



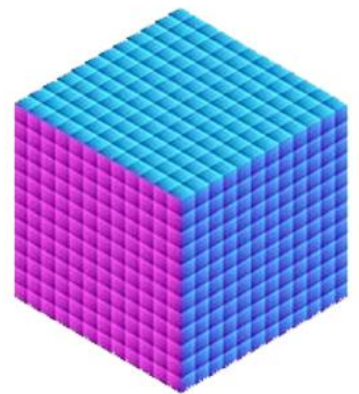
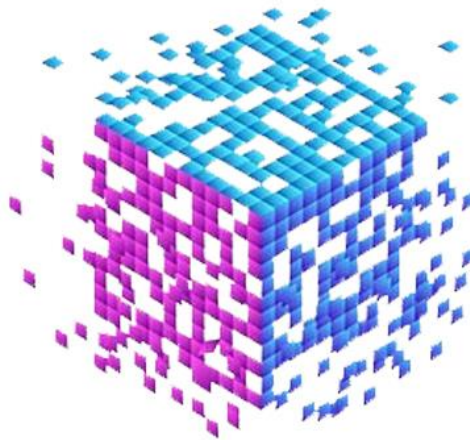
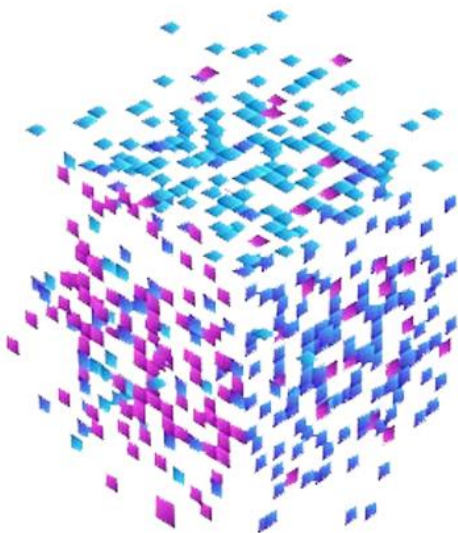
BIG DATA



ANALYTICS



DECISIONS



یکی از نیازهای ضروری سازمان‌های دارای داده‌های حجیم و متنوع، **شناخت دقیق از منابع داده‌ای خود و در اختیار داشتن توانایی مدیریت ذخیره، انتقال، پردازش و بصری‌سازی آن‌ها** است. همواره بزرگترین مشکل در تیم‌های فنی و متخصصین و دانشمندان علم داده، چگونگی دسترسی به داده‌ها و منابع پردازشی مورد نیاز برای تأمین نیازهای دانشی سازمان می‌باشد. تنظیم ساز و کاری برای نگهداری داده‌ها بر اساس پارامترهای زمانی، تعیین سطح دسترسی افراد به منابع داده‌ای، چگونگی و زمان‌بندی واکنشی داده‌ها، شناسایی دقیق محل بروز خطا در زمان انتقال، دریافت شناخت دقیق از کیفیت داده به منظور انتخاب مدل‌های مناسب یادگیری ماشینی و دهه‌ها نیازمندی دیگر می‌تواند سازمان را به سمت ایجاد یک بستر اختصاصی و چارچوب مناسب برای بر عهده گرفتن این وظایف حرکت دهد.

نرم‌افزار **دلتابان**، به منظور پاسخ‌دهی به این مسائل و ایجاد **یک بستر واحد، جامع و مناسب** به منظور استانداردسازی تمامی فعالیت‌های حوزه کلان داده در سازمان، توسط تیم برنامه‌نویسی و کلان داده شرکت سیکاس طراحی و تولید شده است که در ادامه به کلیات آن اشاره شده است.



BIG DATA

ویژگی‌های عمومی نرم افزار دلتابان

• رابط کاربری تحت وب

امروزه رابط تحت وب به عنوان ساده‌ترین و منعطف‌ترین روش برقراری ارتباط با نرم‌افزارها، مورد استفاده و استقبال اکثریت کاربران قرار گرفته است. این نوع رابط کاربری، به راحتی قابل تغییر یا توسعه بوده و کاربر قادر است در هر زمان و از هر کجا، به آن دسترسی داشته باشد. رابط تحت وب این نرم‌افزار، قابلیت‌های مناسبی را برای راهبری این سامانه و مدیریت داده‌های کلان سازمان، ارائه می‌کند.

• امنیت بالا

با توجه به این موضوع که این سامانه به داده‌های اطلاعاتی سایر سامانه‌های مستقر در سازمان دسترسی داشته یا دارد و حتی ممکن است بسته به نیاز، بخشی از داده‌ها نیز در بانک اطلاعاتی این سامانه ذخیره شده باشد، لذا مراقبت از داده‌ها و امنیت بالای این سامانه، در اولویت اول این شرکت بوده و سعی شده تمام مسائل امنیتی در این رابطه، مدنظر قرار گیرد. جدای از این حساسیت‌های، این نرم‌افزار توسط یکی از شرکت‌های امنیتی مطرح کشور، به روش Gray Box و Black Box تست امنیتی شده و **موفق به کسب مجوز امنیتی با درجه امنیت خیلی خوب** شده است.

• ارائه مداوم تغییرات و قابلیت‌های جدید (پویایی بالا)

با توجه به اینکه تیم توسعه این نرم‌افزار به واسطه حوزه فعالیت و تخصص خود، همواره در لبه تکنولوژی کلان داده و داده کاوی در حرکت هستند، لذا دائما در حال ارتقای این محصول و افزودن قابلیت‌های جدید به آن هستند. به طوریکه به صورت معمول، در هر سال حداقل در دو مرحله، این نرم‌افزار به روز رسانی شده و نسخه‌های جدید آن منتشر می‌شود.

• قابلیت اجرا بر روی کلاستر جهت دستیابی به سرعت و توان پردازشی بالا

طبیعتا نرم‌افزاری که قرار است حجم بالایی از داده‌ها را مدیریت و آنالیز نماید، باید ضمن اینکه از هسته قدرتمندی برخوردار باشد، از قابلیت اجرای توزیعی نیز برخوردار باشد. لذا نرم‌افزار **دلتابان**، با این نگاه طراحی و پیاده سازی شده و قادر است بر روی سیستم‌های توزیع شده (کلاستر) اجرا شود.

• قابلیت کار با انواع داده

این نرم‌افزار قادر است تا با **انواع داده‌های ساخت یافته** (نظیر انواع بانک‌های اطلاعاتی رابطه‌ای)، **داده‌های غیر ساخت یافته** (نظیر تصاویر، داده‌های شبکه‌های اجتماعی مانند تلگرام و توییتر)، داده‌های استریمینگ و سایر داده‌های اطلاعاتی موجود در سازمان‌ها نظیر فایل‌های اکسل، اکسس و ... کار کند.

• تجميع و تلفيق داده‌های مختلف

این نرم افزار قادر است تا داده‌های موجود در سازمان از انواع مختلف را با هم تركيب نموده و با يك پرس و جو، پاسخ مورد نیاز مسئله را از این تركيب داده‌ای استخراج نماید.

• استخراج دانش

اجرای انواع مدل‌های یادگیری ماشین جهت استخراج دانش از دل داده‌های خام و بلا استفاده انباشته شده در سازمان که می‌تواند برای تصمیم سازی و تصمیم گیری‌های کلان سازمان، مورد استفاده قرار گیرد.

• امکانات و بخش‌های متعدد متناسب با هر نوع درخواست و برای هر نوع سازمانی

همانگونه که در ادامه خواهید دید، این نرم‌افزار از بخش‌ها و ماژول‌های متنوعی برخوردار می‌باشد به طوریکه آن را قادر ساخته تا برای هرگونه نیازی از سازمان، راهکار مناسبی جهت پاسخ به آن نیاز داشته باشد.

• ارائه رابط نرم‌افزاری API

در این نرم‌افزار، قابلیت‌هایی برای ارائه خدمات به شرکت‌های استارت‌آپی یا سایر سامانه‌های نرم‌افزاری موجود در سازمان و یا سایر سازمان‌های دیگر، به صورت ارائه رابط نرم‌افزاری به شکل API و به شکل کاملاً امن، پیش‌بینی شده است.



• معماری مفهومی سیستم

معماری نرم‌افزار **دلتابان** به شکل زیر و در چهار بخش دیده شده است. بخش اصلی این نرم‌افزار که قلب این سامانه می‌باشد، موتور پردازشی **دلتابان** هست که بصورت کامل و بر اساس تجربیات تیم توسعه دهنده محصول و نیاز سنجی به عمل آمده از مشتریان این شرکت، طراحی و پیاده سازی شده است.

برای تولید این نرم‌افزار از زبان‌های برنامه نویسی، تکنولوژی‌ها و پلتفرم‌های مختلفی نظیر Spark، Hadoop، Python و ... استفاده شده است.



ریز قابلیت‌های قسمت‌های مختلف نرم‌افزار **دلتا بان**

• بخش استخراج داده‌ها

- امکان تهیه شناسنامه برای تمامی منابع داده‌ای سازمان
- امکان تعیین کیفیت منابع داده‌ای و مشاهده میزان کیفیت داده‌ها در لحظه
- امکان مشاهده میزان پیشرفت سازمان در کیفیت بخشی و پاک‌سازی منابع داده‌ای کثیف
- امکان ایجاد و ذخیره Connection string های (CS) مختلف اعم از SQL(JDBC), Stream, Rest API, Documents, Folders, Twitter , ...
- امکان اتصال به CS ساخته شده و دریافت Schema کامل آن بانک و جداول مربوطه و استخراج ساختار آن‌ها
- امکان Local Run یا Force Run کردن هر کدام از موجودیت‌های ذخیره شده. حالت اول اجراها فقط بر روی داده‌های استخراج شده‌ای که در DataLake ذخیره شده‌اند انجام می‌شود و حالت دوم مربوط به اجرای مجدد و استخراج مجدد داده‌های اصلی و اعمال قوانین روی آن‌ها و پاک‌سازی و ساخت مجدد مکعب‌ها می‌باشد.
- دریافت کلیه داده‌های Schema استخراج شده شامل فیلدها، نوع و اندازه آن‌ها، توضیحات، مقادیر (کمینه، بیشینه، میانه، میانگین و ...) فیلدها و رکوردها و همچنین رکوردهای تکراری، داده‌های Missing و ... و ذخیره در Cassandra
- امکان تولید نمونه فرم جمع‌آوری خصوصیات بانک اطلاعاتی با داده‌های پیش‌فرض جهت تکمیل توسط اپراتور و ذخیره آن
- امکان درج توضیحات توسط اپراتور برای فرم استخراج شده، امکان تعیین درجه اهمیت فیلدها، میزان محرمانگی داده، حریم خصوصی، امکان انتشار، حجم داده‌های درون جدول، نرم رشد داده‌ها و ...
- امکان تعیین انواع پیشنهادی پاک‌سازی روی داده‌ها و فیلدها از قبیل عدم رعایت همسانی واحد زمانی، عدم رعایت همسانی واحد وزن، عدم رعایت واحد طولی، عدم رعایت واحد پولی، عدم رعایت همسانی سایر پارامترها، قرار نداشتن مقادیر در محدوده مجاز، عدم رعایت یکتایی مقادیر و ...



• بخش ساخت ETL/ELT و CUBE های داده‌ای

- شناسایی بعد داده‌ها از نظر ویژگی، تعداد مقادیر، مجموعه داده‌ها
- امکان ایجاد کپی از جداول و Data Frame های استخراج شده
- امکان ساخت Cube های داده به صورت ویزاردی به نحوی که بتوان DF های مورد نظر را انتخاب کرده، اطلاعات Metadata هر کدام را ملاحظه نموده، برای حالات مختلف پیش‌پردازش مثل تصمیم‌گیری درباره داده‌های پرت، ترکیب ویژگی‌ها و ... (امکان نوشتن Query یا انتخاب از قابلیت‌های موجود، این قابلیت‌ها توسط متخصص داده‌ای که با سامانه کار می‌کند قابل انتخاب است) و سپس شناسایی فیلدهایی که باید از روی آن‌ها Cube ساخته شود و Join شوند و در نهایت DF جدید ساخته شده (Cube) در SparkDeltaLake ذخیره شود
- امکان اخذ انواع گزارش‌های گرافیکی از DF ها و Cube ها اعم از تعداد رکورد، سایز روی دیسک، زمان تولید، زمان خواندن و اجرا، ابعاد، تعداد مکعب‌هایی که از DF ها استفاده می‌کنند و ...
- امکان مشاهده گرافیکی میزان نقص داده‌های جداول مبتنی بر نقص رکورد یا فیلد به نحوی که با تعیین یکسری معیار استاندارد بتوان میزان نقص هر جدول را محاسبه کرده و نمایش داد.
- امکان ثبت UC های شناسایی‌شده و همچنین تعیین این که کدام DF استخراج شده در این UC استفاده می‌شود.
- امکان تعیین سطح دسترسی به منظور تأیید یا رد UC ها یا DF های شناسایی‌شده
- امکان مشاهده میزان نقص داده DF های هر UC به صورت گرافیکی



• بخش تحلیل و آنالیز های ML داده ها

امکان مشاهده Sample داده از DF ها یا Cube ها

امکان کار با Meta Data یا آمارهای کلان می توان مقادیر زیر را محاسبه کرد:

• **پسامد داده Frequency:** Mode یا Percentile

• **مکان:** میانگین یا میانه

• **پراکندگی:** واریانس، انحراف از معیار، واریانس یا انحراف معیار مبتنی بر میانه (MAD-AAD-Interquartile Range)

امکان تولید وب سرویس و API بر روی موجودیت های سیستم اعم از DF و جداول، مکعب های داده ای و ...

امکان ساخت DF های جدید از DF های استخراج شده در قالب یک Cube یا Data Mart

اتصال موتور کد نویسی به موتور Machine Learning اسپارک برای ایجاد قابلیت استفاده برنامه نویسی در محیط ادمین از الگوریتم های مربوطه

امکان کد نویسی برای استخراج منابع داده ای از شبکه های اجتماعی بر اساس کلمه کلیدی

امکان ذخیره سازی نتیجه یک کد نوشته شده در قالب DF به نحوی که قابل استفاده برای سایرین باشد.

امکان افزودن یک مدل یادگیری ماشینی از قبل نوشته شده در قالب پلاگین به سیستم و استفاده از آن در همه بخش های نرم افزار

امکان تولید API از مدل های یادگیری ماشینی افزوده شده به سیستم



• بخش یادگیری ماشینی (در حال توسعه)

- امکان اجرای الگوریتم‌های کاهش ابعادی هم‌چون PCA و SVD روی DF ها یا Cube ها و ذخیره DF یا Cube جدید
- امکان تعریف گرافیکال گردش کار استخراج و اجرای مدل‌های یادگیری ماشینی و تولید خودکار کدها و تغییر آنها
- امکان تولید API از نتایج مدل‌های یادگیری ماشینی تولید شده به صورت گرافیکال
- امکان تولید انواع نمودارهای گرافیکی روی DF ها و Cube ها (آمارهای کلان، مصورسازی)
- امکان اتصال یک Code Editor به این ابزار به منظور کدنویسی الگوریتم‌های مورد نیاز (مانند Jupyter Notebook)
- امکان استفاده از تمامی کتابخانه‌های رایج برنامه‌نویسی کلان داده در حوزه یادگیری ماشینی اعم از MLlib-Keras-Tensorflow و ...
- امکان کدنویسی در WebIDEA طراحی‌شده در این چارچوب به نحوی که کاربران به سایر ابزارهای کدنویسی نیازی نداشته باشند.
- امکان اتصال به Git برای کنترل ورژن ها و اتصال به Jenkins به عنوان CI به منظور مدیریت Deployment

• بخش بصری سازی داده ها

• مصورسازی: تولید نمودارهای زیر برای همه DF ها و Cube ها:

- هیستوگرام
- نمودارهای جعبه‌ای Box Plot
- نمودارهای پراکندگی Scatter
- نمودارهای بدنه‌ای Contour
- نمودارهای ماتریسی Matrix،
- نمودارهای مختصات موازی Parallel Coordinator
- نمودارهای ستاره‌ای
- صورتک‌های چرنف Chertoff Faces
- امکان اتصال ابزارهای بصری سازی مانند Tableau، Superset و ...
- امکان کار با انواع داده‌های GIS و جغرافیایی (در حال توسعه)



• بخش مدیریتی

- امکان تعریف کلاسترهای مجازی و ایجاد قابلیت تعیین این که کد نوشته شده روی کدام کلاستر اجرا شود
- امکان تعیین سطوح دسترسی مختلف برای کاربران:
- تعریف کاربر، انتساب UC ها به کاربران، تعیین سطح دسترسی کاربر به فیلدها یا تعداد رکوردهای مشخصی از هر DF، تعیین سطح دسترسی کاربر به کلاستر مجازی
- امکان تعریف زمانبندی اتصالات به منابع داده‌ای و بازه‌های زمانی به روزرسانی دریاچه داده
- امکان مشاهده گزارش نتیجه واکنشی داده‌ها از نظر سلامت اجرا یا خطا در زمان واکنشی اطلاعات
- امکان ایجاد یک پایپ لاین از مرحله استخراج داده تا لحظه بصری‌سازی
- امکان قابلیت مانیتورینگ داده در تمام دوره حیات آن (LineAge)
- امکان در اختیار داشتن تمامی مدل‌های رایج یادگیری ماشینی برای جلوگیری از کدنویسی مجدد
- امکان استفاده مجدد کاربران از نتایج کارهای قبلی و سایرین



نسخه‌های نرم افزار دلتا بان

نام پکیج	Standard	Enterprise	Rocket	Jet
قابلیت	★★★★★	★★★★★	★★★★★	★★★★★
ماژول پایه (Core Module) مبتنی بر اسپارک	✓	✓	✓	✓
داشبورد تولید گزارشات آماری	✓	✓	✓	✓
امکان نوشتن توابع پاکسازی جدید (User Defined Function)		✓	✓	✓
امکان تولید Open API برای دسترسی به داده های انبار داده (دسترسی به CUBE ها) با قابلیت تعیین دسترسی کاربران به API ها		✓	✓	✓
امکان اتصال به داده های Streaming و دریافت آنها		✓	✓	✓
امکان اتصال به محیط کد نویسی مبتنی بر وب Web IDEA			✓	✓
امکان ساخت کلاسترهای مجازی و ایجاد دسترسی برای گروه های برنامه نویسی			✓	✓
امکان تعریف قوانین سازمانی دسترسی به داده ها برای اجرای Data Governance			✓	✓
امکان اتصال به مدل های یادگیری ماشینی و اجرای آنها				✓
فعال سازی قابلیت اجرای نرم افزار بر روی GPU به جای CPU				✓
محدودیت تعداد فید برای اتصال به منابع داده ای بیرونی	۱۰	۲۰	نامحدود	نامحدود
نصب و راه اندازی و آموزش نرم افزار	✓	✓	✓	✓

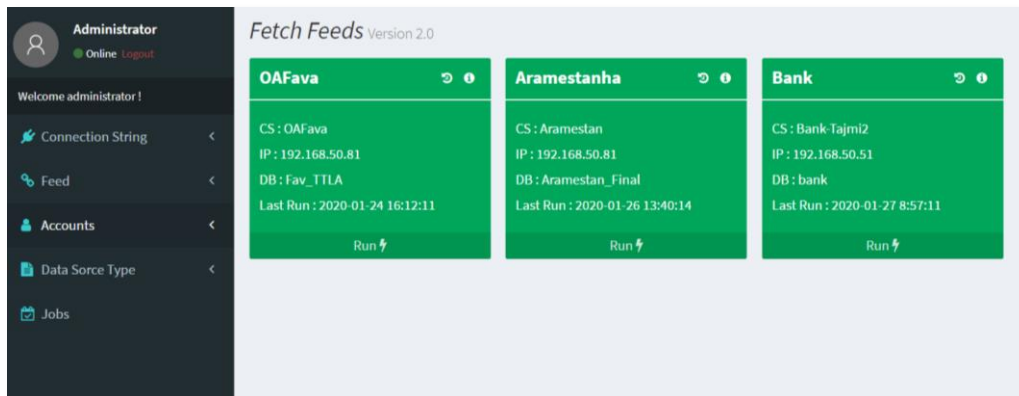


امکانات ماژول پایه یا Core Module :

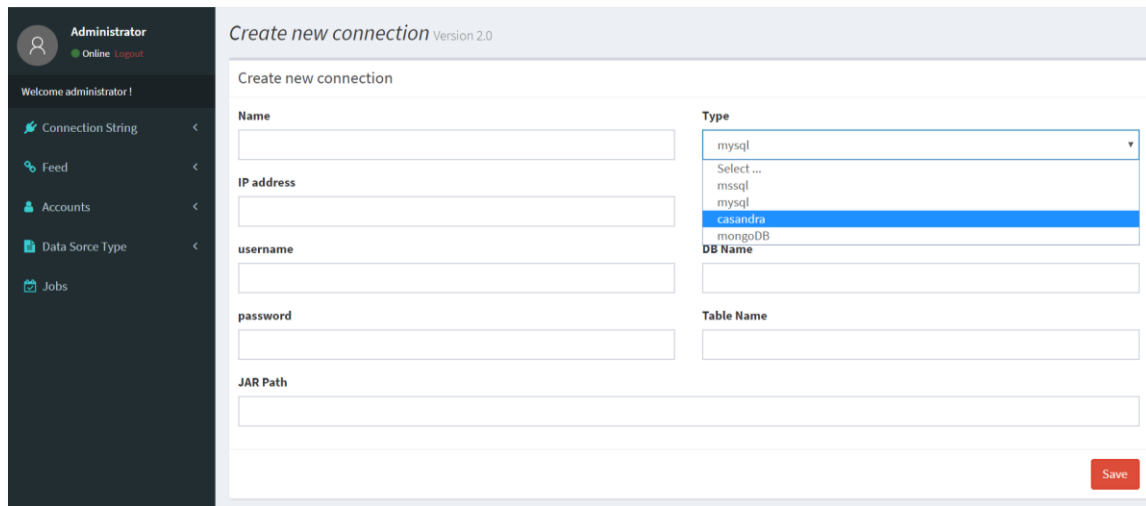
- قابلیت اتصال به بانک‌های اطلاعاتی مختلف اعم از MSSQL, MySQL, Postgres, Oracle و فایل‌ها
- داده‌ای با فرمت‌های استاندارد
- امکان ساخت دریاچه داده Data Lake
- امکان شناسنامه دار کردن بانک‌ها و جداول خوانده شده
- امکان ثبت شناسنامه داده‌ها
- امکان انجام عملیات Profiling بانک‌های اطلاعاتی
- امکان ثبت Business Glossary
- امکان تعریف انواع کوئری‌ها برای Extract از بانک‌های اطلاعاتی موجود
- امکان استفاده از توابع پاکسازی نوشته شده
- امکان ذخیره سازی داده‌های خوانده شده در قالب Data Frame در دریاچه داده
- امکان انتقال داده‌ها به حافظه رم
- امکان نوشتن SQL برای دستکاری جداول
- امکان ساخت Cube‌ها
- دسترسی به داشبورد مدیریتی سیستم
- امکان تعریف کاربران سیستم
- امکان مشاهده وضعیت استفاده از منابع سخت افزاری
- امکان تعریف زمانبندی برای به روزرسانی داده‌ها



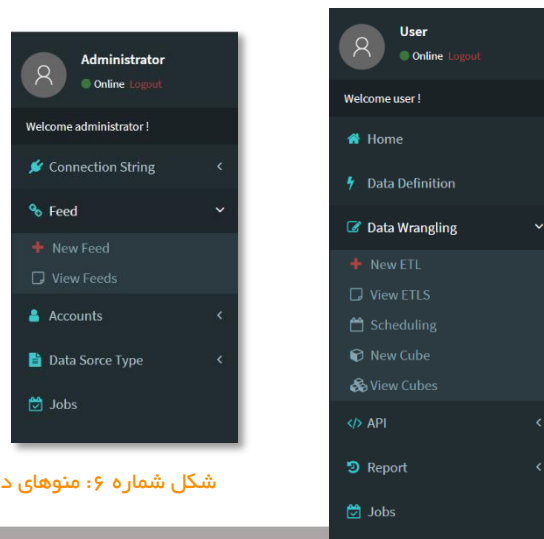
بخش‌هایی از رابط کاربری نرم‌افزار



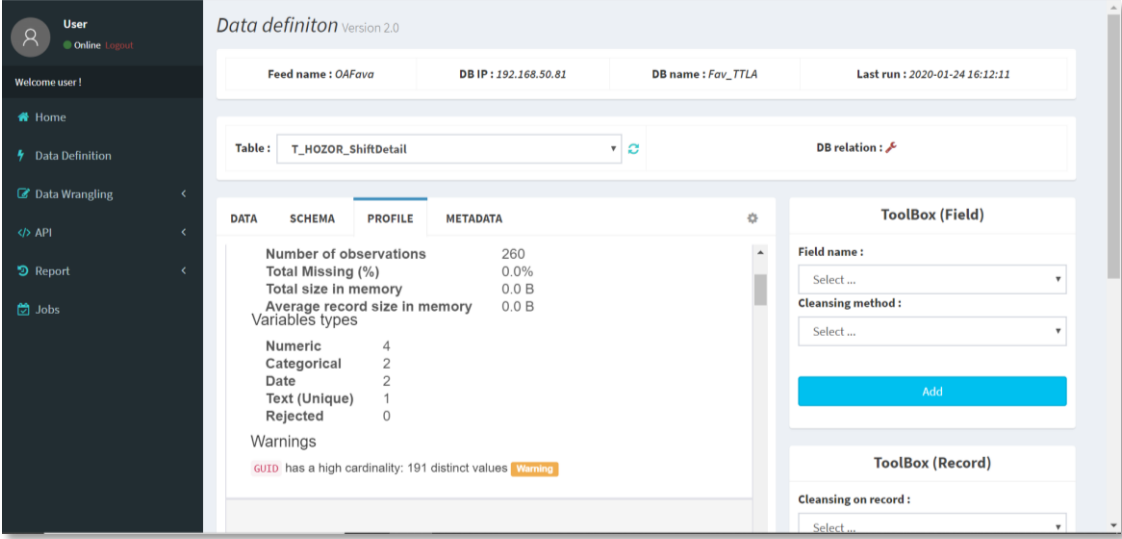
شکل شماره ۴: وضعیت دیتابیس‌های فید شده جهت استخراج متادیتا



شکل شماره ۵: تعریف اتصال جدید به یک دیتابیس جهت انجام عملیات Profiling



شکل شماره ۶: منوهای دو حالت کاربری Administrator و User



Data definition Version 2.0

Feed name : OAFava DB IP : 192.168.50.81 DB name : Fav_TTLA Last run : 2020-01-24 16:12:11

Table : T_HOZOR_ShiftDetail DB relation :

PROFILE

Number of observations	260
Total Missing (%)	0.0%
Total size in memory	0.0 B
Average record size in memory	0.0 B
Variables types	
Numeric	4
Categorical	2
Date	2
Text (Unique)	1
Rejected	0

Warnings

GUID has a high cardinality: 191 distinct values **Warning**

ToolBox (Field)

Field name : Select ...

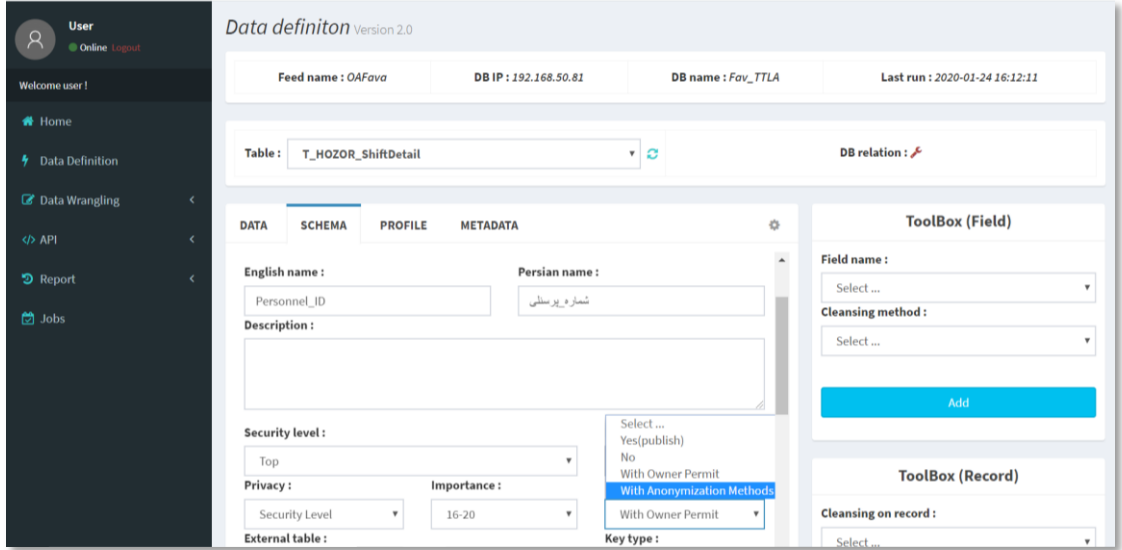
Cleansing method : Select ...

Add

ToolBox (Record)

Cleansing on record : Select ...

شکل شماره ۷: خروجی گزارش Profiling دیتابیس



Data definition Version 2.0

Feed name : OAFava DB IP : 192.168.50.81 DB name : Fav_TTLA Last run : 2020-01-24 16:12:11

Table : T_HOZOR_ShiftDetail DB relation :

SCHEMA

English name : Personnel_ID Persian name : شماره پرسنلی

Description :

Security level : Top

Privacy : Security Level Importance : 16-20

External table : Key type : With Anonymization Methods

ToolBox (Field)

Field name : Select ...

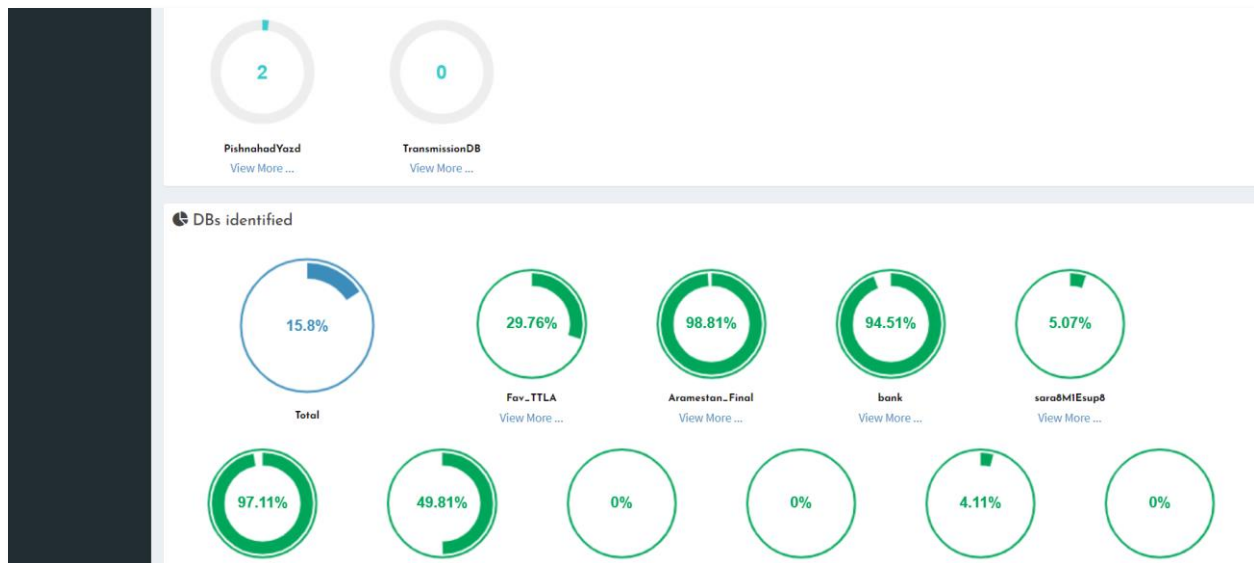
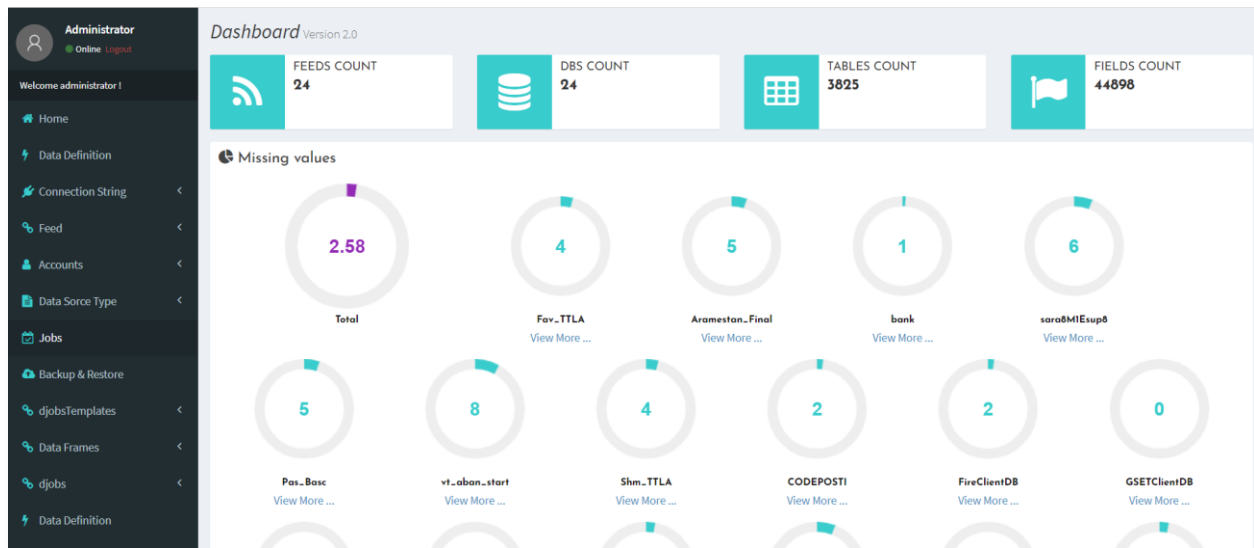
Cleansing method : Select ...

Add

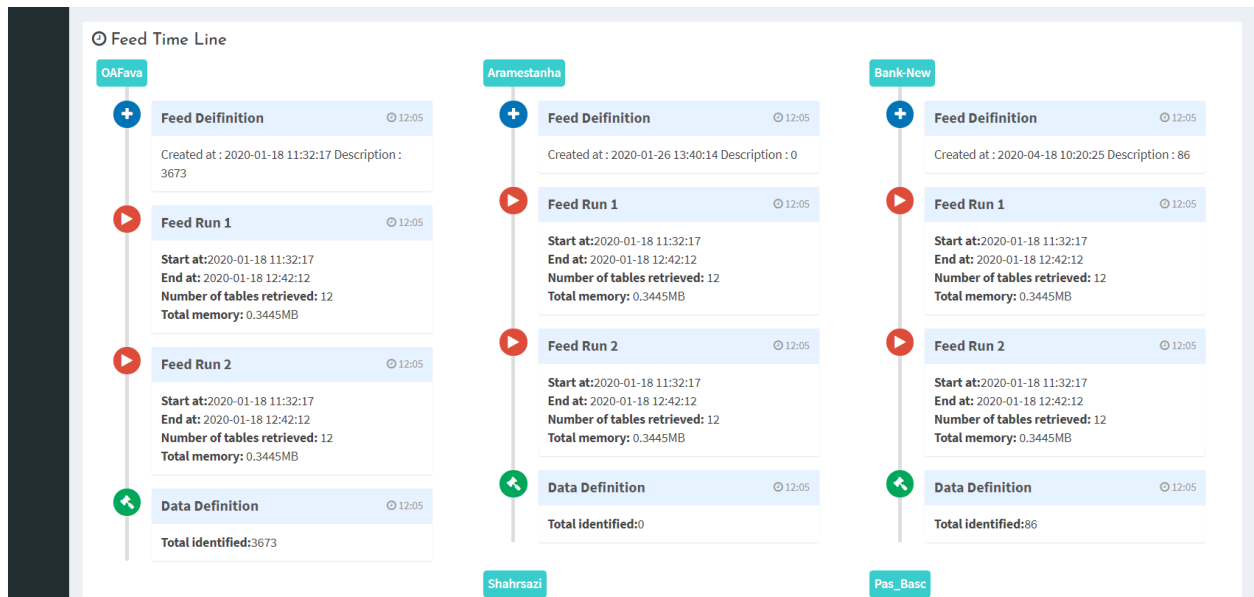
ToolBox (Record)

Cleansing on record : Select ...

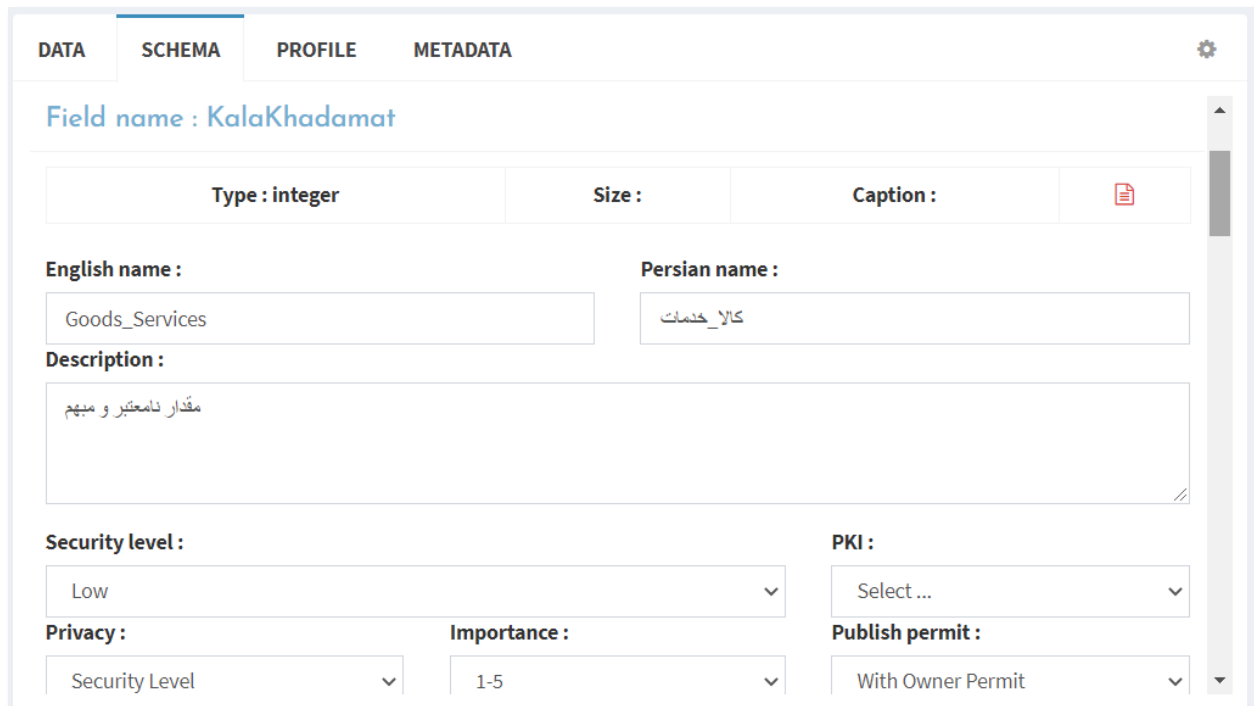
شکل شماره ۸: فرم شناسنده دار کردن جدول اطلاعاتی (Data Definition)



شکل شماره ۹: بخشی از داشبورد



شکل شماره ۱۰ : تایم لاین واکنشی داده ها



Field name : KalaKhadamat

Type : integer	Size :	Caption :
English name : Goods_Services Persian name : کالا_خدمات Description : مقدار نامعتبر و مبهم		
Security level : Low	Importance : 1-5	PKI : Select ...
Privacy : Security Level	Publish permit : With Owner Permit	

شکل شماره ۱۱ : متادیتاها

Administrator
Online [Logout](#)

Welcome administrator!

- Home
- Data Definition
- Connection String
- Feed
- Accounts
- Data Source Type
- Jobs
- Backup & Restore
- djobsTemplates
- Data Frames
- djobs
- Data Definition

Extract Version 2.0

Feed: Bank-New Dataframe: Scheduler: Select ... Security level: Select ... Save to disk: ☐ **RUN**

initializing

Bank-New

- access_control
- avarez
- banks
 - attr_name
 - bank_name
 - bank_code
- channel_type
- codenosazi
- collision
- dlog
- export_delimiter
- export_items
- farabar_avarez_type
- file_contents
- file_info
- nosazi_log
- parsian_contents
- picture

1 Select * from banks where bank_name=

ALTER
AND
AS
ASC
BETWEEN
BY
COUNT
CREATE
DELETE
DESC
DISTINCT
DROP
FROM
GROUP

شکل شماره ۱۲: ETL نویسی

Identified tables Version 2.0

Select feed

Feed: Bank-New

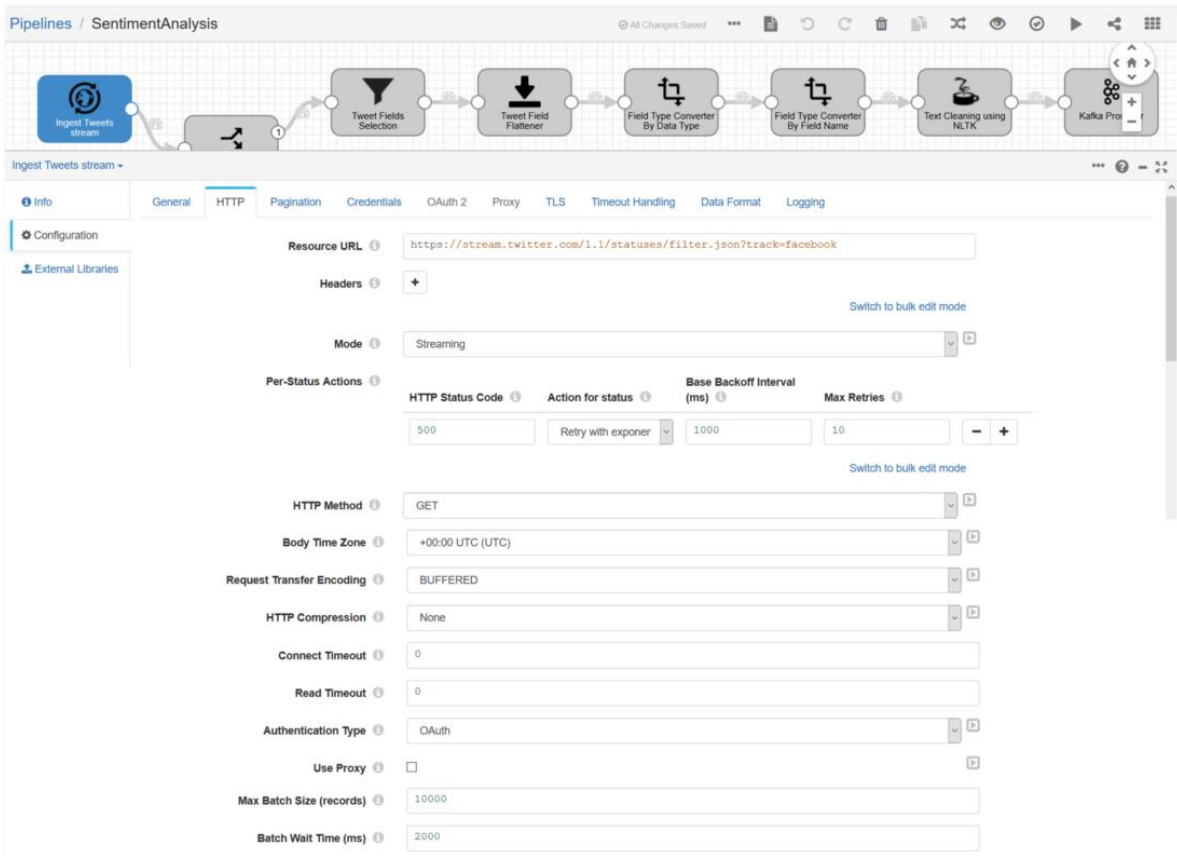
No.	Table	Rows	Fields	weight	ProgressBar	Progress	MetaData	Info
0	access_control	48	5	undefined		100%	✓	i
1	avarez	2	5	undefined		100%	✓	i
2	banks	23	3	undefined		100%	✓	i
3	channel_type	10	3	undefined		100%	✓	i
4	codenosazi	37526	3	undefined		100%	✓	i
5	collision	1974	8	undefined		100%	✓	i
6	dlog	0	5	undefined		0%	✗	i
7	export_delimiter	2	2	undefined		100%	✓	i
8	export_items	7	2	undefined		100%	✓	i
9	farabar_avarez_type	4	3	undefined		100%	✓	i

شکل شماره ۱۳: گزارش

بخش‌هایی از رابط کاربری نسخه پیشرفته محصول



شکل شماره ۱۴: توسعه گرافیکی فرایند اجرای مدل یادگیری ماشینی (نسخه پیشرفته)



Pipelines / SentimentAnalysis

Ingest Tweets stream

Configuration

Resource URL: `https://stream.twitter.com/1.1/statuses/filter.json?track=facebook`

Headers: +

Mode: Streaming

Per-Status Actions

HTTP Status Code	Action for status	Base Backoff Interval (ms)	Max Retries
500	Retry with exponer	1000	10

HTTP Method: GET

Body Time Zone: +00:00 UTC (UTC)

Request Transfer Encoding: BUFFERED

HTTP Compression: None

Connect Timeout: 0

Read Timeout: 0

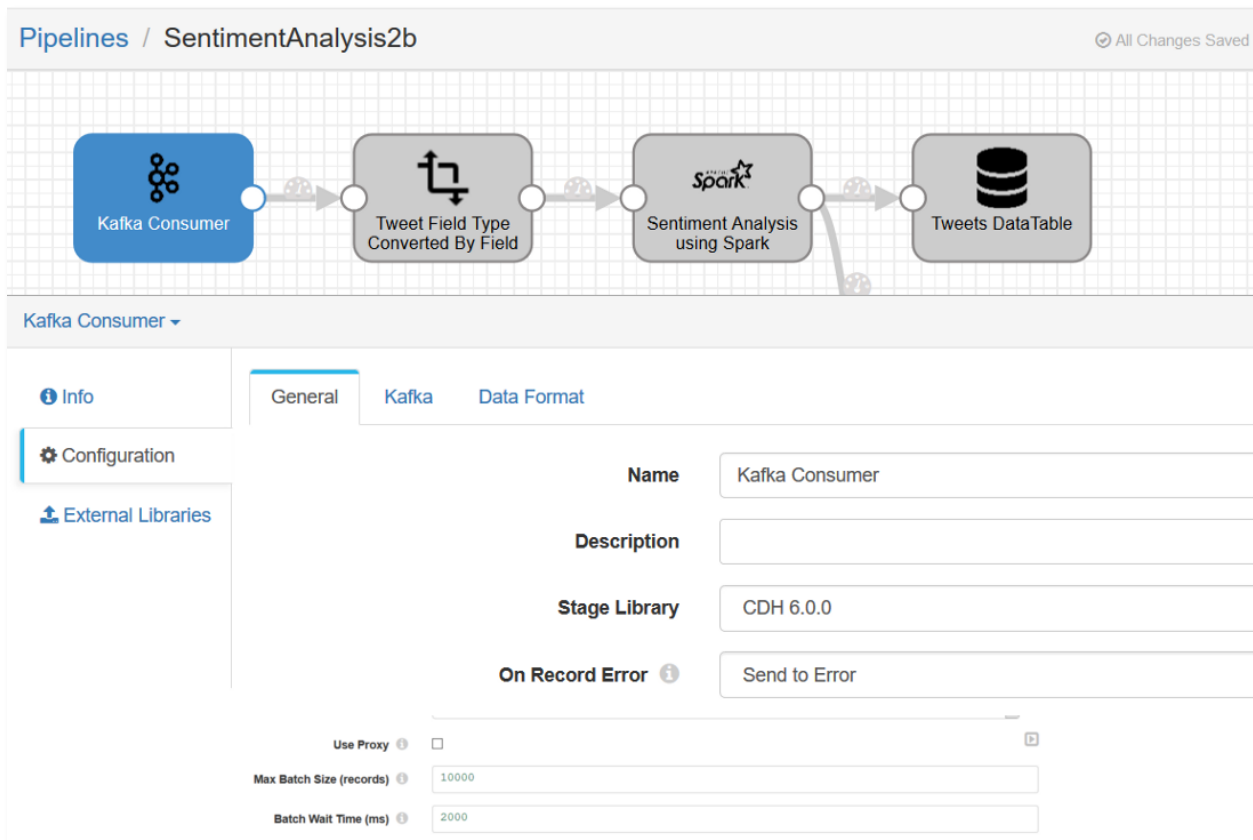
Authentication Type: OAuth

Use Proxy: ☐

Max Batch Size (records): 10000

Batch Wait Time (ms): 2000

شکل شماره ۱۵: توسعه گرافیکی فرایند اجرای مدل یادگیری ماشینی (نسخه پیشرفته)



شکل شماره ۱۶: توسعه گرافیکی فرایند اجرای مدل یادگیری ماشینی (نسخه پیشرفته)

پروژه‌های انجام شده

این سامانه در حال حاضر بصورت کامل در شهرداری یزد نصب شده و از اردیبهشت سال ۱۳۹۹ مورد بهره برداری قرار گرفته است. در این پروژه مراحل زیر بصورت کامل و مبتنی بر این نرم افزار، در کنار کار کارشناسان داده این شرکت، انجام گرفته است:

- شناسایی منابع داده‌ای (بانک‌های اطلاعاتی) بر اساس خوداظهاری کارفرما
- اتصال به بانک‌های اطلاعاتی و استخراج فراداده (**Meta Data**) آنها (از طریق نرم افزار)
- شناسنامه دار کردن جداول و فیلدهای استخراج شده از بانک‌های اطلاعاتی (کارشناسی)
- انجام عملیات **ETL** روی داده‌ها (از طریق نرم افزار)
- ساخت انبار داده مبتنی بر پلتفرم اسپارک و هادوپ (از طریق نرم افزار)