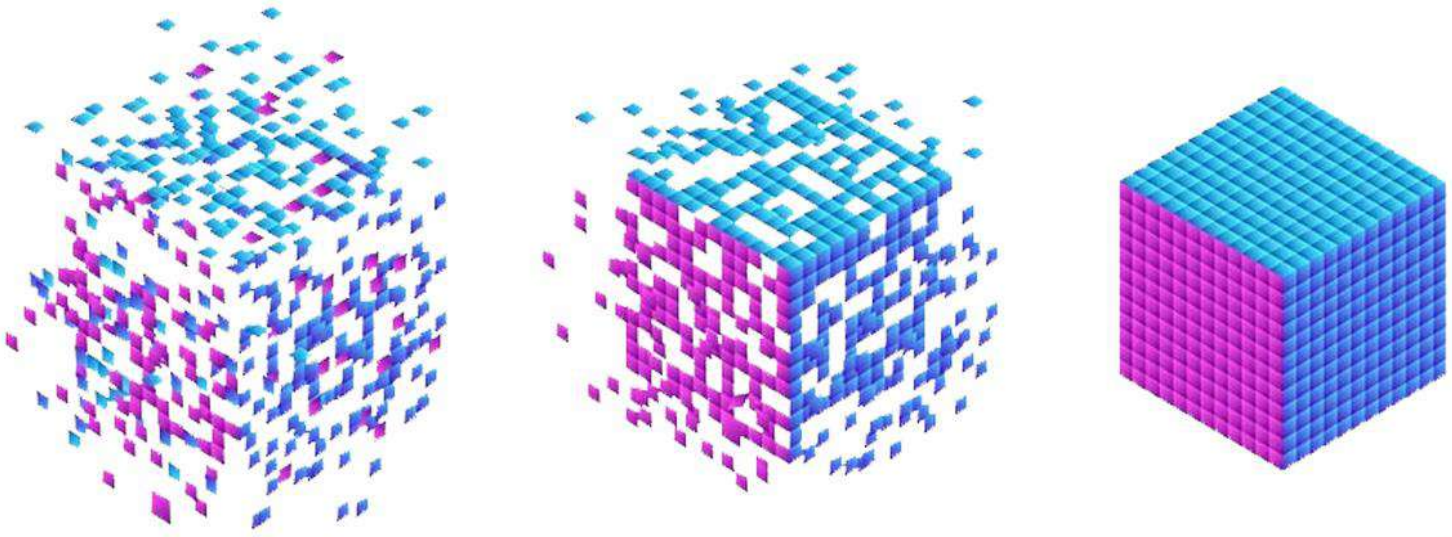
سکاس
SECAAS

سامانه نرم‌افزاری

دلتا بان
DELTA BAAN

فریم‌ورک انباره داده و مدیریت چرخه حیات داده‌ها

BIG DATA ▶ ANALYTICS ▶ DECISIONS



bigdata.secaas.ir

یکی از نیازهای ضـروری سـازمان‌های دارای داده‌های حجیم و متنوع، **شـناخت دقیق از منابع داده‌ای خود** و **در اختیار داشـتن توانایی مدیریت ذخیره، انتقال، پردازش و بصـری‌سـازی آن‌ها** اسـت. همواره بزرگترین مشـکل در تیم‌های فنی و متخصـصـین و دانشـمندان علم داده، چگونگی دسـترسـی به داده‌ها و منابع پردازشـی مورد نیاز برای تأمین نیازهای دانشـی سـازمان می‌باشـد. تنظیم سـاز و کاری برای نگهداری داده‌ها بر اسـاس پارامترهای زمانی، تعیین سـطح دسترسی افراد به منابع داده‌ای، چگونگی و زمان‌بندی واکشـی داده‌ها، شـناسـایی دقیق محل بروز خطا در زمان انتقال، دریافت شـناخت دقیق از کیفیت داده به منظور انتخاب مدل‌های مناسب یادگیری ماشینی و ده‌ها نیازمندی دیگر می‌تواند سـازمان را به سـمت ایجاد یک بسـتر اختصـاصـی و چارچوب مناسـب برای بر عهده گرفتن این وظایف حرکت دهد.

نرم‌افزار ***دلتـابـان***، به منظور پاسخ‌دهی به این مسائل و ایجاد **یک بستر واحد، جامع و مناسب** به منظور اسـتانداردسـازی تمامی فعالیت‌های حوزه داده شامل مدیریت دریاچه داده، کاتالوگ داده، واژه‌نامه کسـب و کار، توسـعه مدل‌های یادگیری ماشینی (گرافیکی و کدنویسی)، ساخت انواع ETL ها، حاکمیت داده‌ها در سـازمان، توسـط تیم برنامه‌نویسـی و کلان داده شـرکت سـیکاس طراحی و تولید شده است که در ادامه به کلیات آن اشاره شده است.

ویژگی‌های عمومی نرم‌افزار 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 حساسیت‌ها، این نرم‌افزار توسط یکی از شرکت‌های امنیتی مطرح کشور، به روش Black Box و Gray 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 داده‌کاوی در حرکت هستند، لذا دائما در حال ارتقاء این محصول و افزودن قابلیت‌های جدید به آن هستند. به طوری‌که به صورت معمول، در هر سال حداقل در دو مرحله، این نرم‌افزار به روزرسانی 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 پیاده‌سازی شده و قادر است بر روی سیستم‌های توزیع شده (کلاستر) اجرا شود. ماشین‌های محاسباتی کلاستر می‌توانند مبتنی بر کوبرنتیز یا به صورت Bare Metal باشند.

- **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.

- **تجمیع و تلفیق داده‌های مختلف**

این نرم‌افزار قادر است تا داده‌های موجود در سازمان از انواع مختلف را با هم ترکیب نموده و با یک پرس و جو، پاسخ مورد نیاز مسئله را از این ترکیب داده‌ای استخراج نماید.

- **استخراج دانش**

اجرای انواع مدل‌های یادگیری ماشین جهت استخراج دانش از دل داده‌های خام و بلااستفاده انباشته شده در سازمان که می‌تواند برای تصمیم‌سازی و تصمیم‌گیری‌های 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**

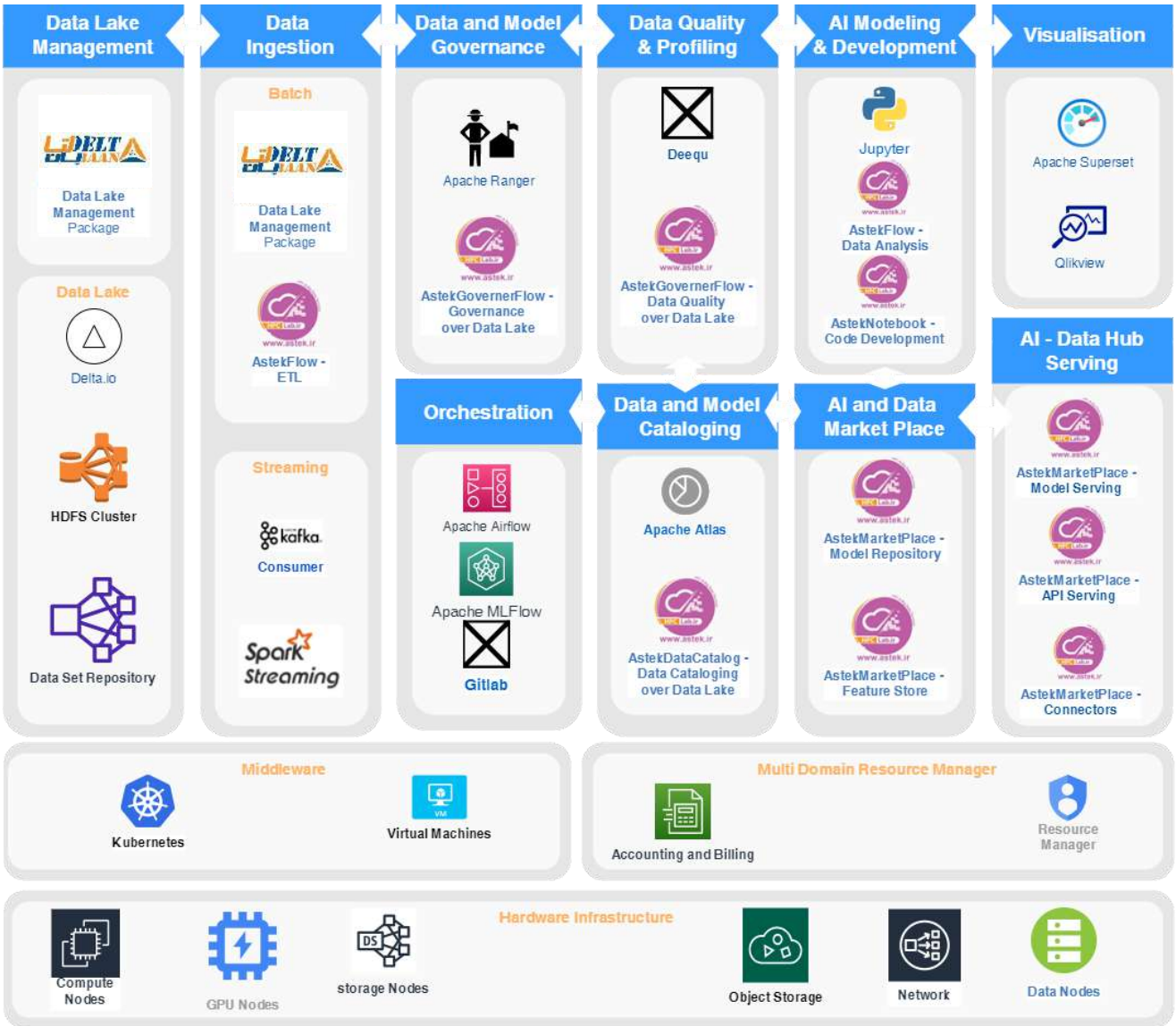
در این نرم‌افزار، قابلیت‌هایی برای ارائه خدمات به شرکت‌های استارتاپی یا سایر سامانه‌های نرم‌افزاری موجود در سازمان و یا سایر سازمان‌های دیگر، به صورت ارائه رابط نرم‌افزاری به شکل **API** و به شکل کاملا امن، پیش‌بینی شده است.

شکل شماره ۱: مشخصات دلتابان

- **معماری مفهومی سیستم**

معماری نرم‌افزار **دلتـــابـــان** به شکل زیر و در چهار بخش دیده شده است. بخش اصلی این نرم‌افزار که قلب این سامانه می‌باشد، موتور پردازشی **دلتـــابـان** هست که به صورت کامل و بر اساس تجربیات تیم توسعه‌دهنده محصول و نیازسنجی به عمل آمده از مشتریان این شرکت، طراحی و پیاده‌سازی شده است.

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



شکل شماره ۲: معماری سامانه

ریز قابلیت‌های قسمت‌های مختلف نرم‌افزار دلتابان

- **امکانات پایه**

- رابط کاربری تحت وب
- دارای گواهی امنیت افتا
- قابلیت اجرا بر روی سیستم‌های توزیع‌شده (Kubernetes Clusters)
- دارای خروجی‌های API با فرمت JSON جهت تبادل و استفاده از داده‌ها
- امکان تعریف گروه‌های کاربری و نقش‌ها و اعمال دسترسی‌های تعریف شده در قالب Policy ها به گروه‌های کاربری
- امکان تعریف کاربران سیستم و تعیین دسترسی آنان بر اساس Policy های تعریف‌شده برای DF ها و API ها
- مشاهده وضعیت استفاده از منابع اختصاص‌یافته به سامانه
- امکان دسترسی به متادیتاهای ثبت‌شده در سامانه با کمک قابلیت ساخت Connection String و اتصال به بانک‌های اطلاعاتی سامانه
- سازگاری کامل با تاریخ شمسی و دارابودن ابزار تقویم شمسی
- امکان دریافت ساختارهای اطلاعاتی جهت ایجاد Data Frame
- امکان نصب سامانه در سیستم‌عامل‌های Linux و Windows و امکان استفاده از مرورگرهای IE-Chrome-Firefox برای واسط کاربری
- Dockerize بودن نرم‌افزار
- ارائه لیست ماژول‌ها و زیرسیستم‌های موجود در ابزار به منظور انطباق شرح خدمات با آن‌ها
- ارائه برنامه زمان‌بندی نصب و بهره‌برداری و عملیاتی‌سازی و آموزش هر یک از آن‌ها

- **بخش مدیریتی**

- امکان تعریف کلاستر مجازی برای نصب سامانه بر بستر آن (مبتنی بر Kubernetes)
- امکان تعیین سطوح دسترسی برای کاربران به نحوی که کاربران در قالب گروه‌هایی تعریف شده و به هر گروه بتوان Policy مربوط به DF-AOI انتساب داد
- تعریف کاربر، تعیین سطح دسترسی کاربر به فیلدها یا تعداد رکوردهای مشخصی از هر Data Frame (بر اساس تعریف یک Where Clause در بخش تعریف Policy ها)
- ثبت لاگ کاربران با امکان تهیه گزارش از آن‌ها
- امکان تعریف زمان‌بندی اتصالات به منابع داده‌ای و بازه‌های زمانی به‌روزرسانی دریاچه داده
- امکان مشاهده گزارش نتیجه واکشی داده‌ها از نظر سلامت اجرا یا خطا در زمان واکشی اطلاعات
- امکان ایجاد یک Pipeline از مرحله استخراج داده تا قبل از بصری‌سازی
- امکان قابلیت مانیتورینگ داده در تمام دوره حیات آن در محیط داخلی Data Lake به نحوی که مشخص باشد که Data Frame مربوطه چه زمانی ایجاد شده، چه افرادی با چه Query هایی به آن دسترسی داشته‌اند، چه API هایی با چه مشخصه‌هایی از آن دیتافریم را فراخوانی کرده‌اند
- امکان Transform کردن (ترکیب چندین Data Frame، تفکیک آن) Data Frame های موجود در سامانه و ساخت Data Frame جدید از آن‌ها به نحوی که Query مربوط به ساخت آن Data Frame قابل مشاهده بوده و سایر کاربران نیز بتوانند به محتوای Data Frame جدید ساخته‌شده دسترسی داشته باشند.
- امکان تفکیک Data Frame های سامانه در سه لایه Raw Data- Transformed Data- Feature Store به نحوی که کاربر ارشد سامانه بتواند Data Frame ها را در هر یک از این لایه‌ها قرار دهد. لایه دوم مربوط به Data Frame های جدیدی است که از ترکیب Data Frame های لایه اول و دوم ساخته می‌شوند و لایه سوم حاوی Data Frame هایی خواهد بود که قرار است به کاربران بیرونی به صورت سرویس داده شود و عمدتا داده‌های محرمانه و حریم خصوصی از آن‌ها باید حذف شوند.

- **شناسایی و آماده‌سازی داده‌ها**

کاتالوگ داده

- امکان تهیه شناسنامه از دیتافریم‌های سامانه که از منابع بیرونی استخراج شده‌اند به‌صورت یک منبع داده مجزا و قابل دسترس
- امکان تعریف و دسترسی به منابع داده‌ای با معرفی روش دسترسی به منبع داده‌ای (برای مثال: ایجاد و ذخیره Connection String) مختلف مربوط به پایگاه‌های داده‌ای شامل بانک‌های اطلاعاتی DB2- PostGreSQL- MySQL- Oracle- SQL Server و شبکه اجتماعی تلگرام و فایل فرمت‌های Avro- XML- JSON- CSV- Parquet
- قابلیت ایجاد Data Frame با استفاده از Data Frame های موجود
- قابلیت اعمال دسترسی روی Data Frame ها
- امکان دریافت Schema هر منبع داده‌ای بیرونی در مرحله اتصال شامل نام جداول، نام فیلدها، نوع و سایز فیلدها و استخراج آن‌ها با قابلیت نگهداشت، نمایش و چاپ و جستجوی آن‌ها و اضافه کردن توضیحات در مورد هر قلم داده‌ای و اضافه کردن برچسب‌های اختصاصی (tag) روی هر قلم اطلاعاتی و ثبت داده‌های شناسنامه‌ای مانند نام فارسی و انگلیسی، توضیحات، سطح محرمانگی و حریم خصوصی، درجه اهمیت، روش پیشنهادی پاک‌سازی، امکان انتشار
- قابلیت دریافت داده‌ها به‌صورت Stream و Batch
- امکان دسترسی و تهیه گزارش ویژگی‌های منتسب‌شده به اقلام اطلاعاتی شناسنامه‌دار شده
- امکان استفاده از این نرم‌افزار به عنوان یک Data Hub به نحوی که بتوان از منابع بیرونی داده‌ها را دریافت کرده و در صورت نیاز Transform و پاک‌سازی نموده (با کمک کوئری‌های SQL) و سپس در قالب API های استاتیک و دینامیک به کاربران یا سامانه‌های بیرونی ارسال کرد

- امکان تعریف انواع کوئری‌ها از Data Frame های موجود در سامانه در لایه‌های مختلف
- امکان مشاهده میزان نقص داده‌های دریافت‌شده شامل درصد Missing Value ها به تفکیک جداول یک بانک اطلاعاتی بیرونی، تعداد رکوردها، تعداد فیلدها و نوع آن‌ها، مقدار Cardinality هر فیلد، تعداد رکوردهای تکراری، مقدار Skewness و Kurtosis هر فیلد، میزان Percentile فیلد؛ امکان مشاهده میزان پیشرفت اپراتورها در ثبت متادیتاهای هر Data Frame استخراج‌شده از منابع بیرونی و متادیتاهای فیلدهای آن
- امکان انجام عمل (Profiling) داده‌های بیرونی بدون انتقال داده از منابع داده

کیفی‌سازی داده‌ها

- امکان انجام عملیات نمایه‌سازی (Profiling) بانک‌های اطلاعاتی بیرونی
- امکان اجرای الگوریتم‌های پاک‌سازی/کیفی‌سازی (اعتبارسنجی) روی داده‌ها در قالب کوئری‌های SQL
- امکان تعیین انواع عملیات پیشنهادی جهت 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 به نحوی که برنامه‌نویس SQL در مرحله ساخت دیتافریم‌های لایه یک و دو بتواند این پیشنهادات را مشاهده نماید
- امکان مشاهده میزان کیفیت داده‌ها بر اساس شاخص‌های Missing Value- Percentile Kurtosis- Skewness- Duplicate Value-
- امکان مشاهده میزان پیشرفت سازمان در کیفیت بخشی و پاک‌سازی منابع داده‌های کثیف
- امکان توسعه و اعمال الگوریتم‌های بی‌نام‌سازی روی داده‌ها در قالب کوئری‌های SQL

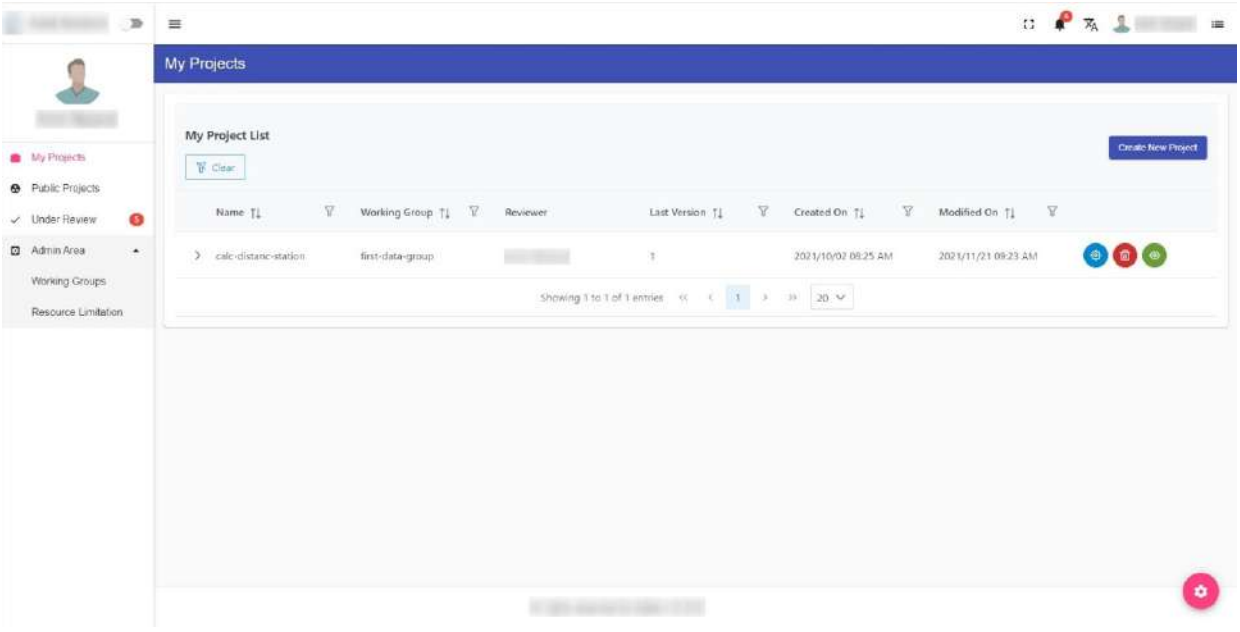
آماده‌سازی و ارائه داده‌ها

- قابلیت ایجاد دریاچه داده به‌صورت مجازی‌سازی داده‌ها با قابلیت profiling روی منبع داده اصلی و یا دریافت داده‌ها از منابع داده‌ای بیرونی
- ذخیره‌سازی داده‌های منتقل شده به دریاچه داده به‌صورت سری زمانی
- قابلیت انتقال تغییرات جدید (از جمله اضافه، ویرایش، حذف) در داده‌های منبع اصلی به دریاچه داده
- قابلیت ارائه داده‌ها به‌صورت مستقیم از منبع داده با در نظر گرفتن سطوح دسترسی تعیین‌شده روی منابع داده مربوطه
- امکان ساخت DF جدید با تجمیع داده‌ها از منابع مختلف شامل منابع خارج از دریاچه داده و نیز منابع داخل دریاچه داده در قالب کوئری‌های SQL
- امکان تعریف زمان‌بندی برای به‌روزرسانی داده‌ها در صورت نیاز
- تهیه کپی از منابع داده‌ای و Data Frame و مشاهده نمونه داده‌ها
- امکان انجام محاسبات SQL پایه روی منابع داده‌ای با استفاده از SPARK و ایجاد منبع داده جدید در دریاچه داده
- امکان اخذ گزارش‌های گرافیکی از Data Frame ها شامل تعداد رکورد، سایز روی دیسک، زمان تولید، زمان خواندن و اجرا
- امکان تولید API بر روی موجودیت‌های شامل DF ها مبتنی بر Policy های ساخته‌شده

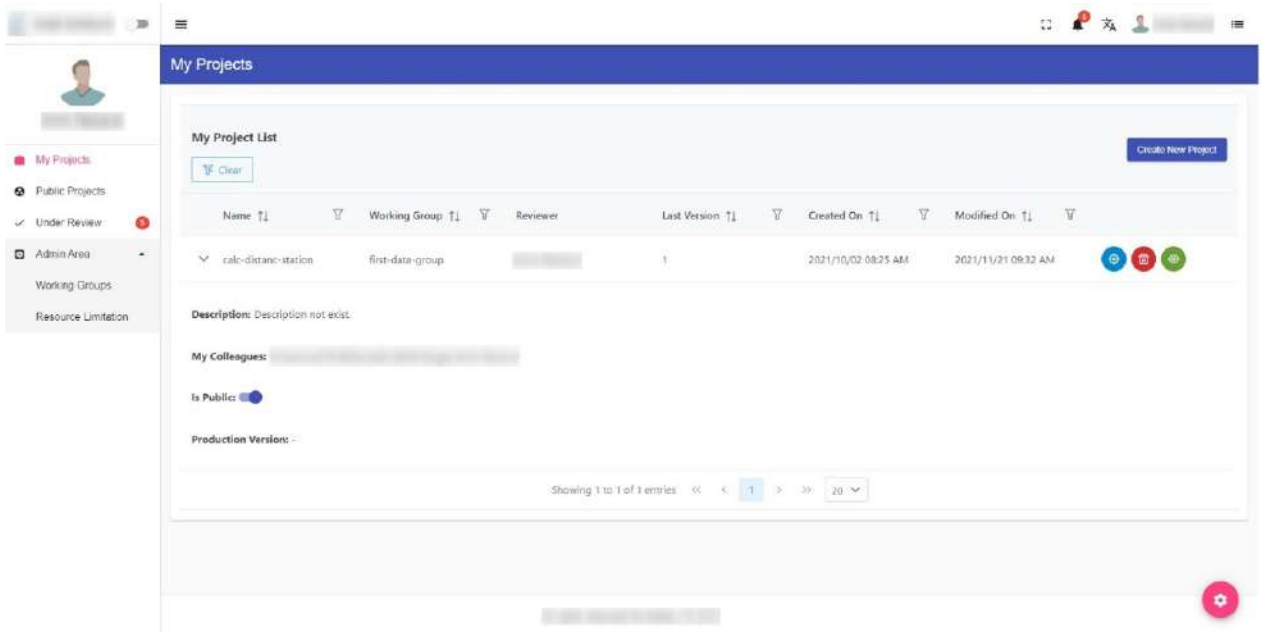
مدیریت پایش موردهای کاربردی (USE CASE Management)

- امکان تعریف Use Case های شناسایی و یا اعلام شده
- امکان ثبت شناسنامه Use Case و پیوست مستندات مربوط به هر Use Case
- ثبت منابع داده‌ای مورد نیاز و منبع داده‌ای موجود به‌صورت چک‌لیست جهت برآورده‌شدن رفع نیاز اعلام شده توسط تیم کارشناسی برای هر Use Case
- امکان جستجو در محتوای Use Case ها
- امکان دریافت گزارش به‌صورت فایل در قالب‌های متداول (word, excel, pdf) از شناسنامه Use Case ها
- امکان تعیین سطح دسترسی گروه‌های کاربری در ثبت، مشاهده و جستجوی Use Case ها

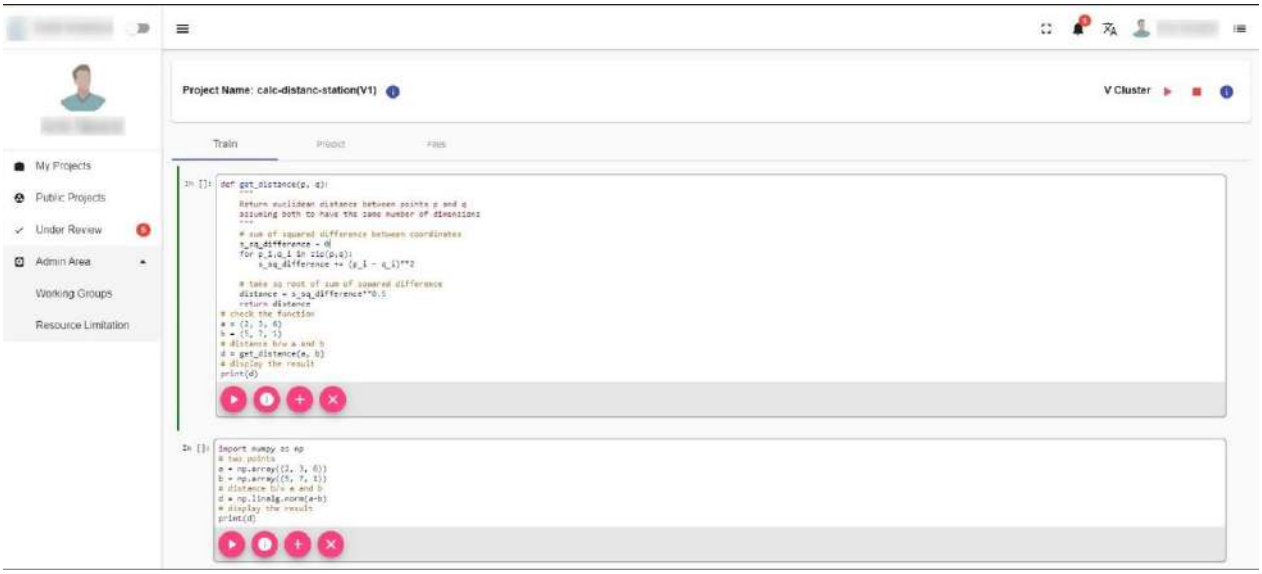
بخش‌هایی از رابط کاربری نرم‌افزار



شکل شماره ۳: محیط کدنویسی

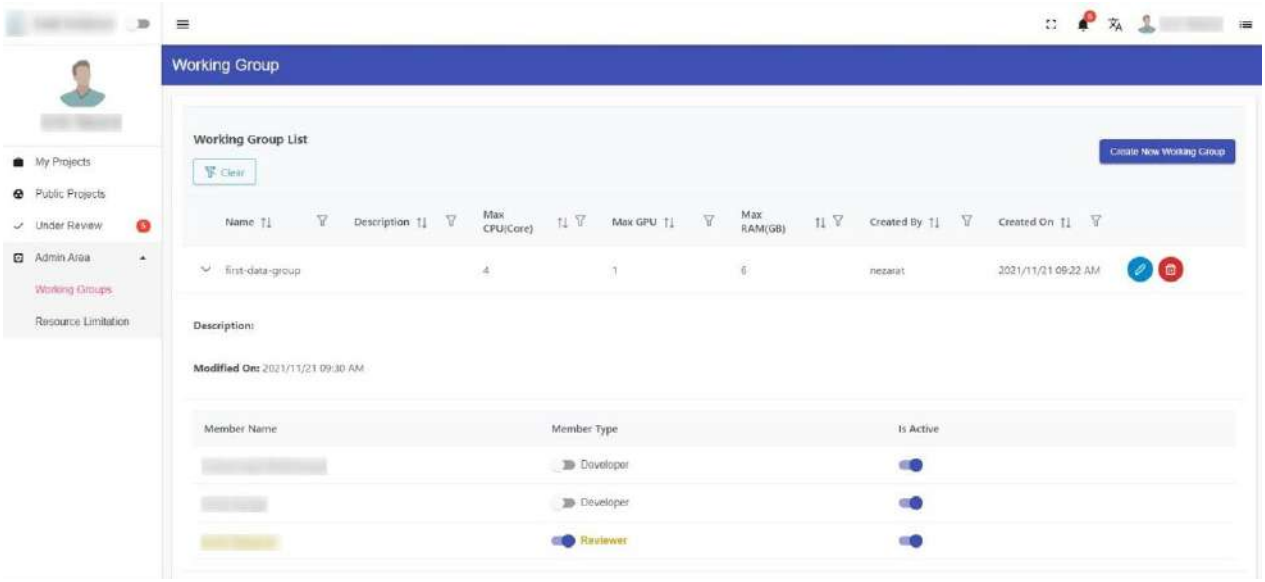


شکل شماره ۴: محیط کدنویسی ۲

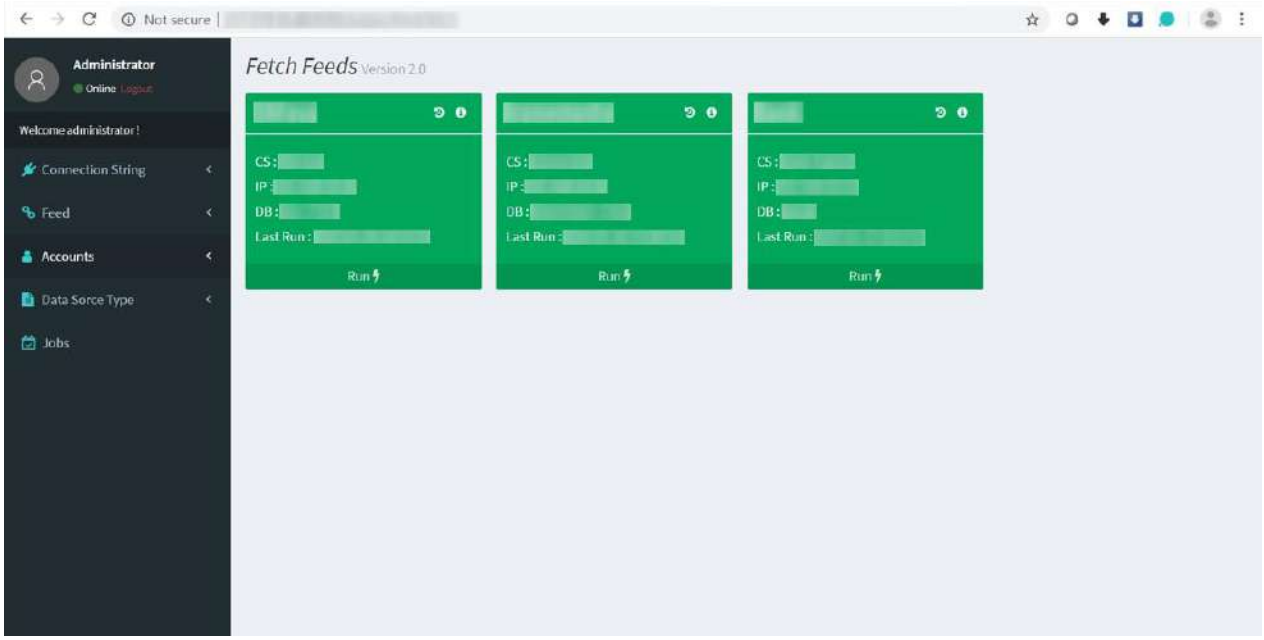


شکل شماره ۵: سلول‌های کدنویسی کاربر

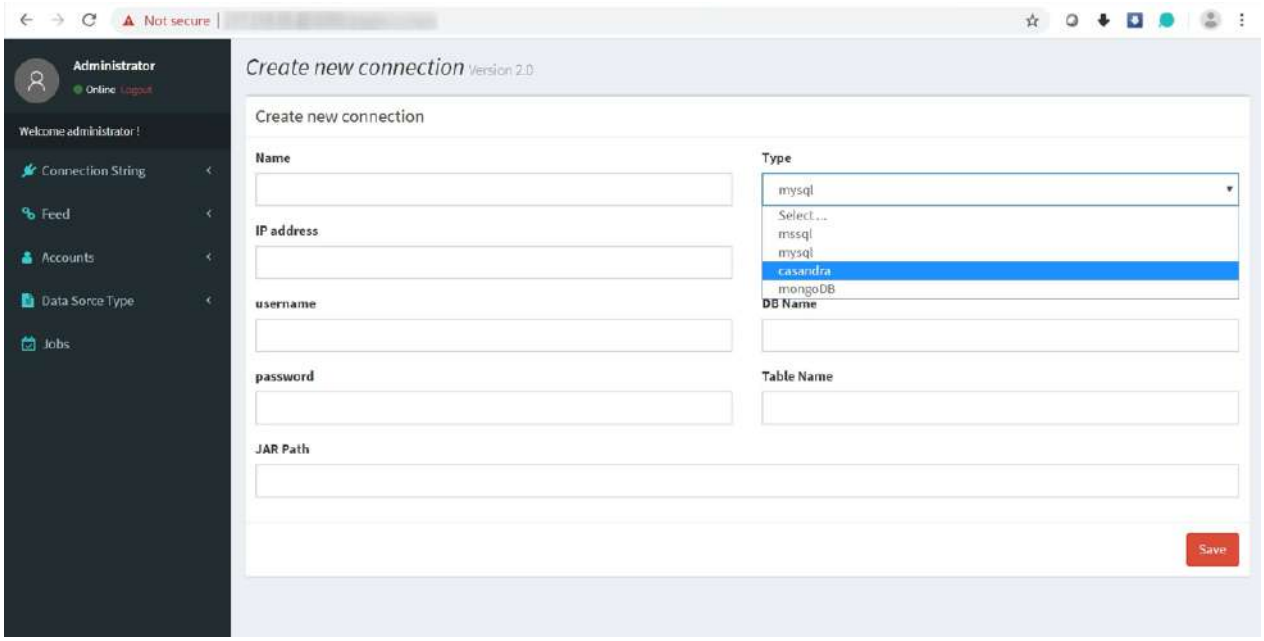
شکل شماره ۶: تعریف منابع محاسباتی تیم‌ها



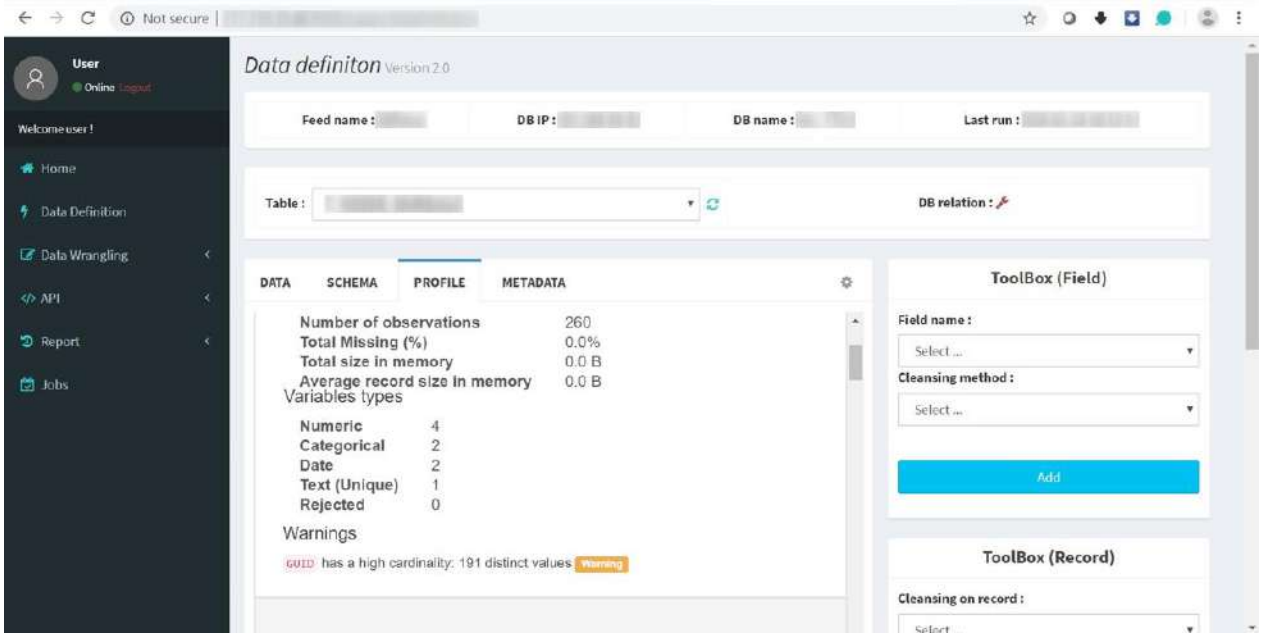
شکل شماره ۷: تعریف گروه‌های کاری



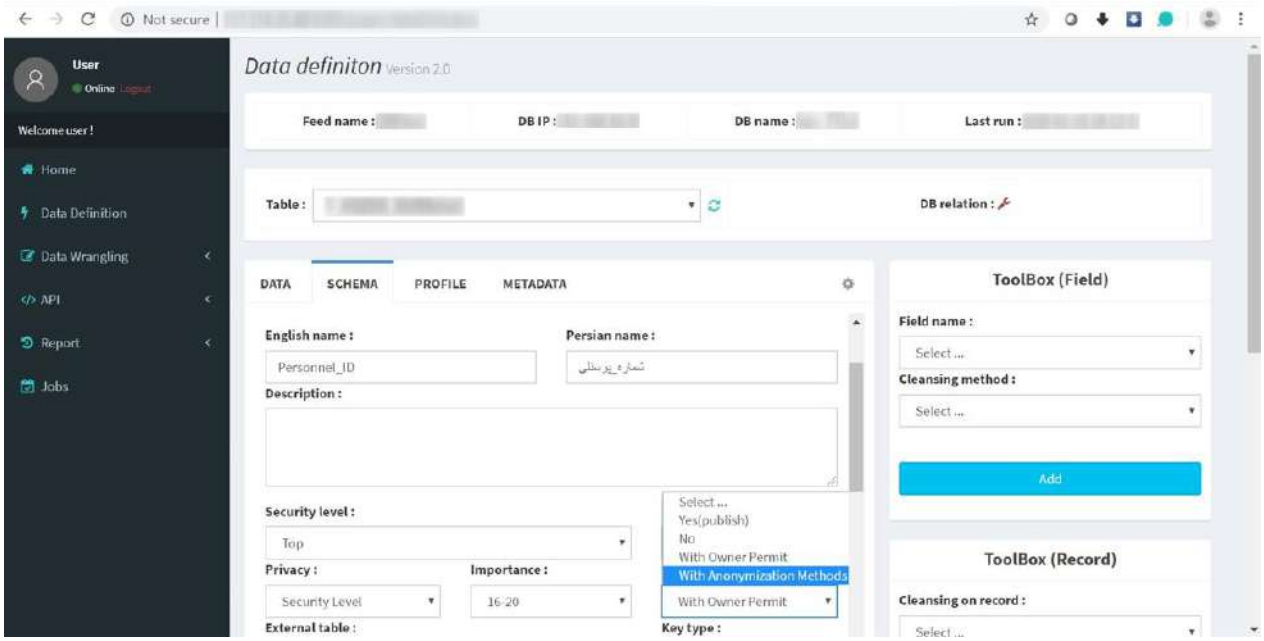
شکل شماره ۸: وضعیت دیتابیس‌های فیدشده جهت استخراج متادیتا



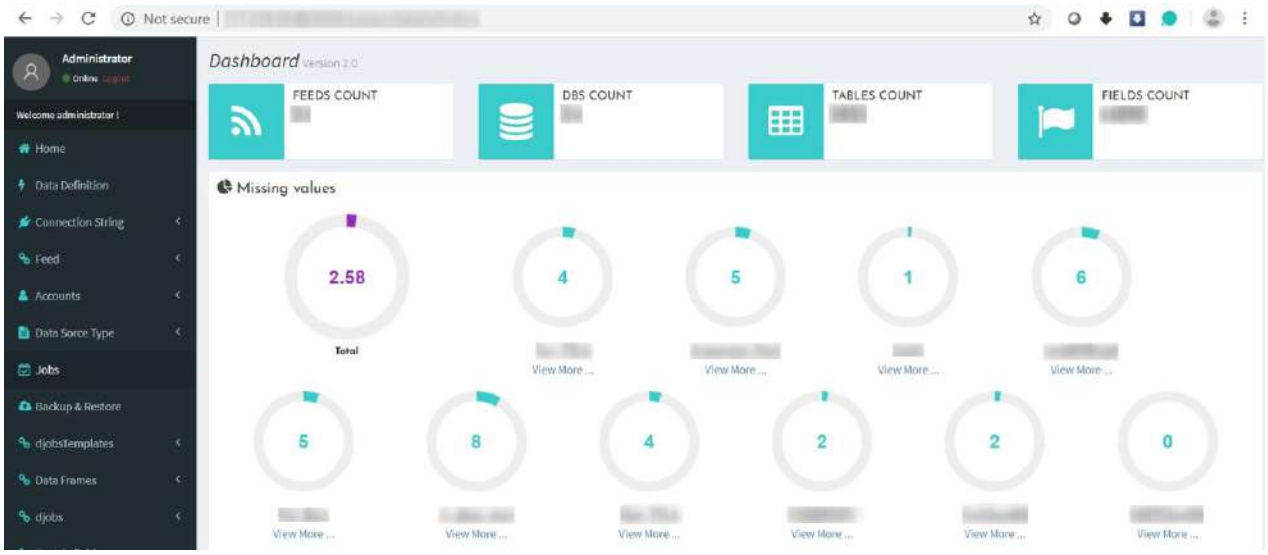
شکل شماره ۹: تعریف اتصال جدید به یک دیتابیس جهت انجام عملیات Profiling



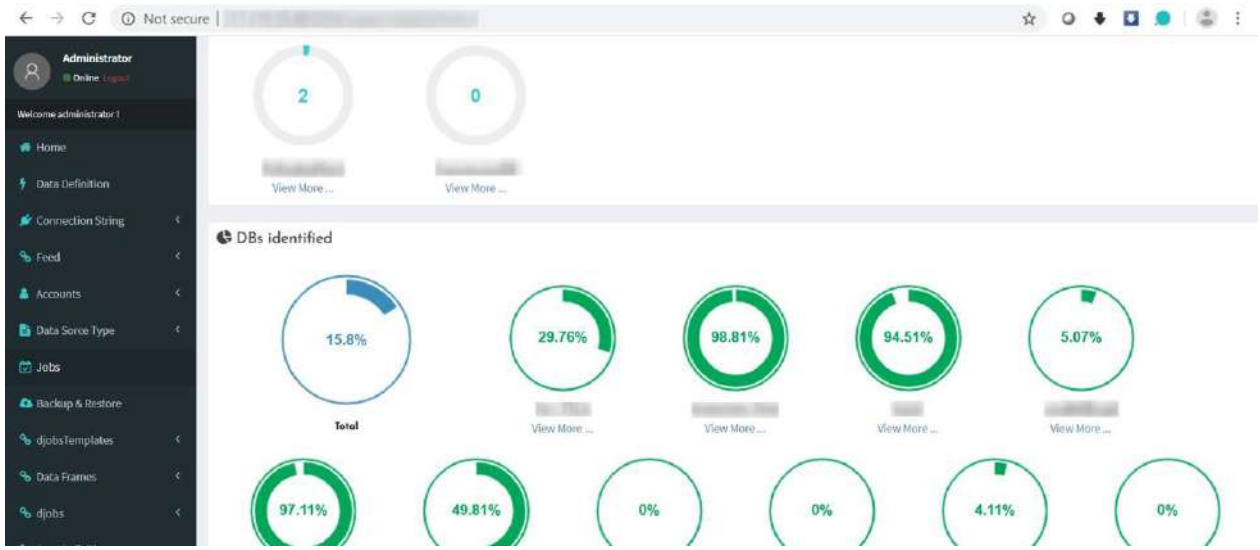
شکل شماره ۱۰: خروجی گزارش Profiling دیتابیس



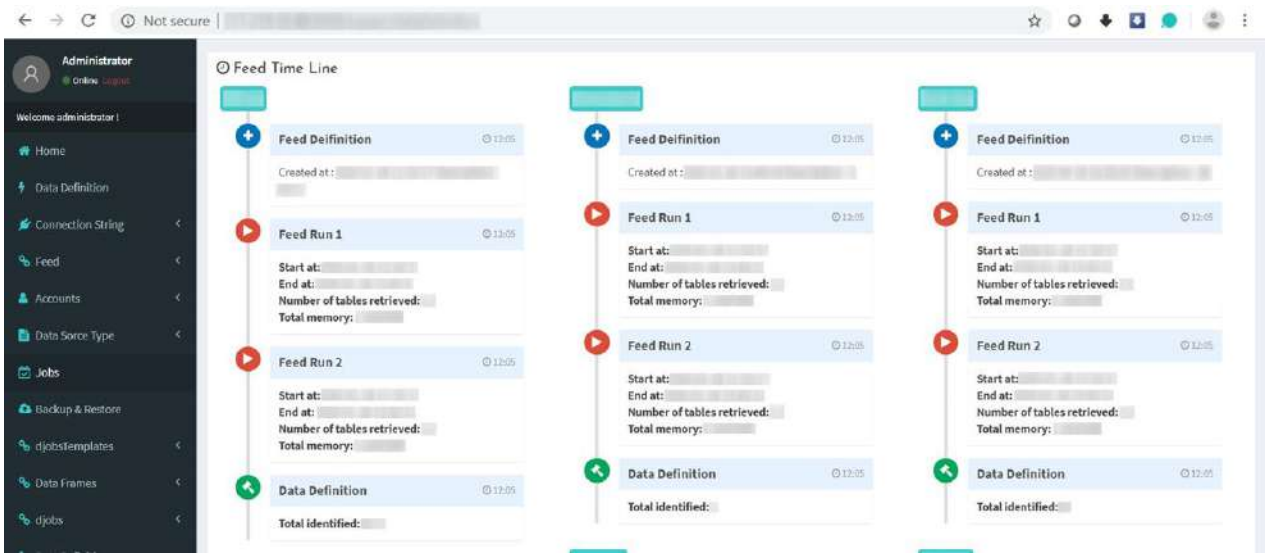
شکل شماره ۱۱: فرم شناسنامه‌دار کردن جدول اطلاعاتی (Data Definition)



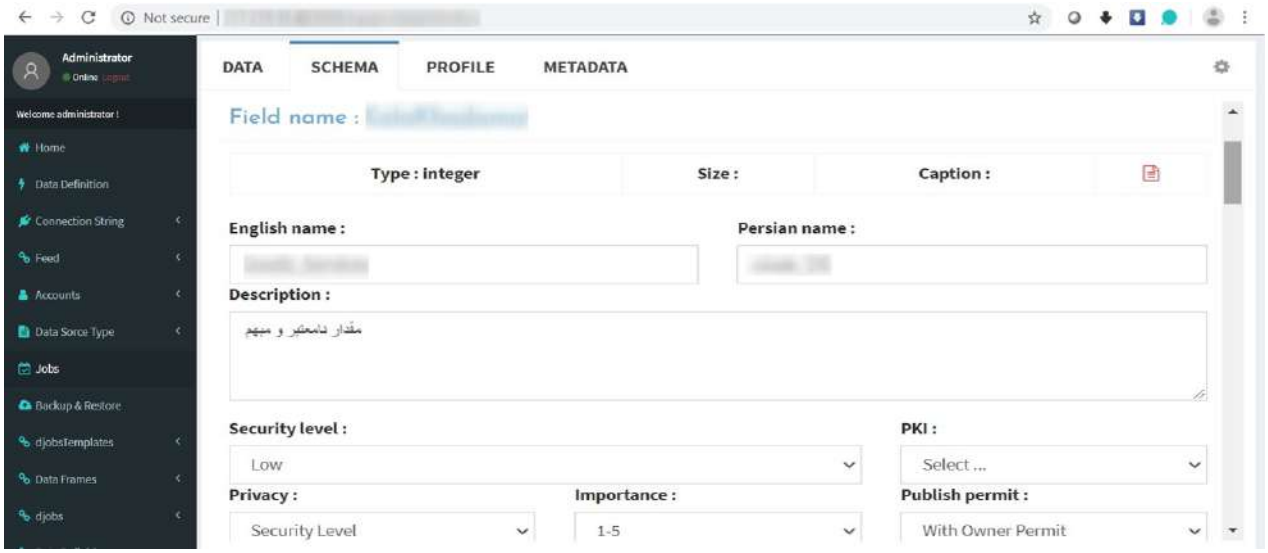
شکل شماره ۱۲: بخشی از داشبورد مدیریتی



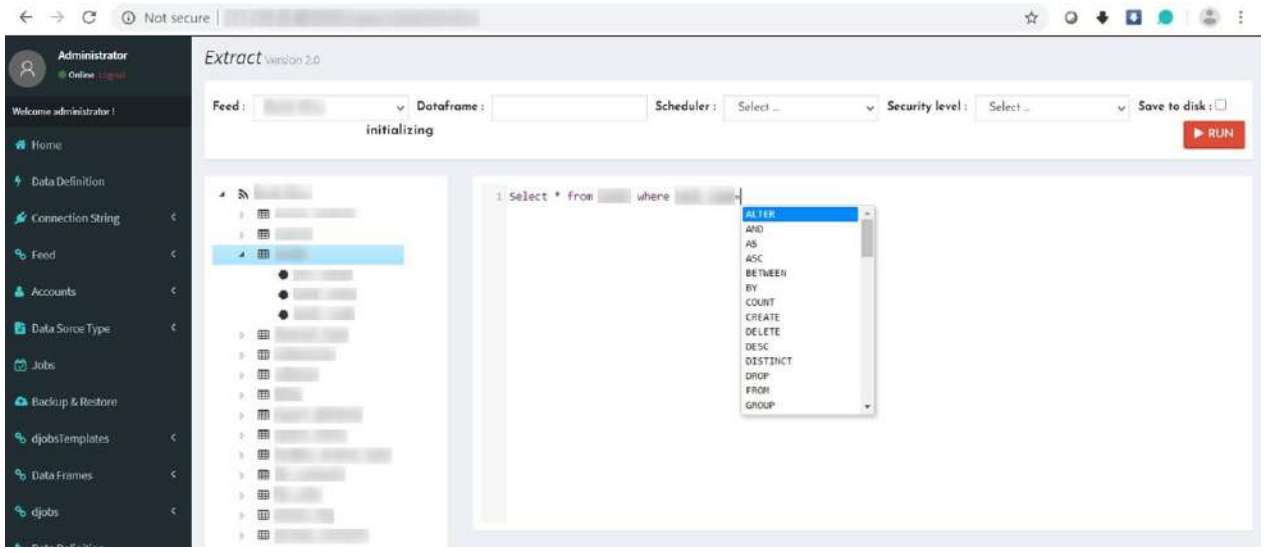
شکل شماره ۱۳ : بخشی از داشبورد مدیریتی ۲



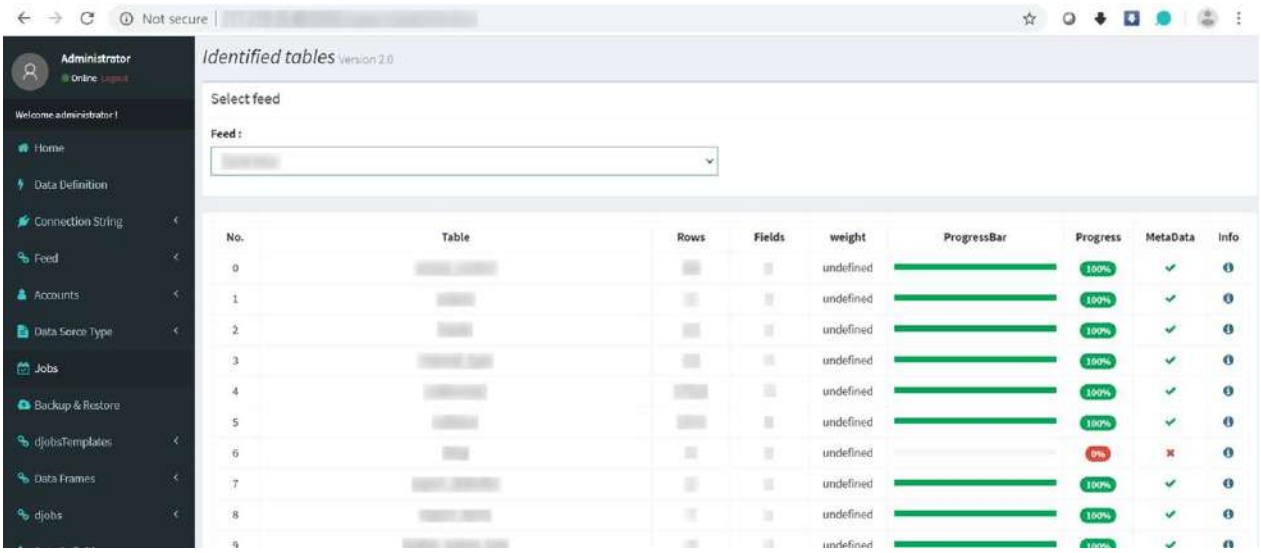
شکل شماره ۱۴ : Timeline واکشی داده‌ها



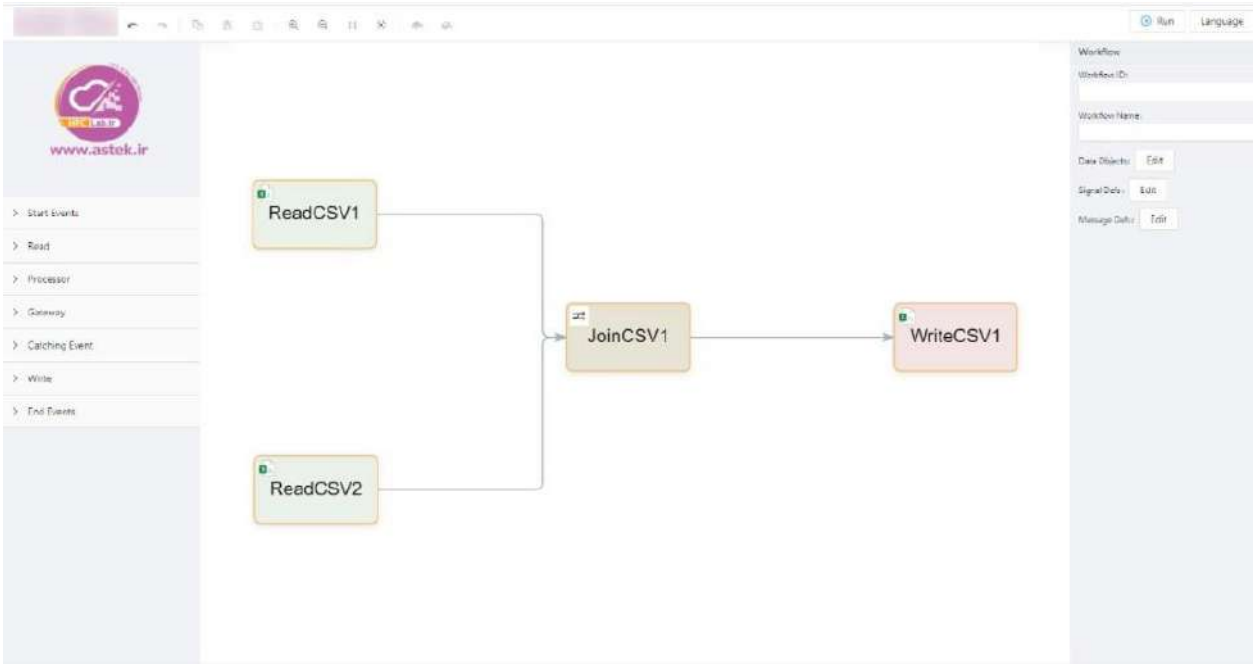
شکل شماره ۱۵ : متادیتاها



شکل شماره ۱۶ : ETL نویسی



شکل شماره ۱۷ : گزارش سیستم



شکل شماره ۱۸ : توسعه گرافیکی فرآیند اجرای مدل یادگیری ماشینی و تعریف ELT

DELTA دلتا BAAN بان

دفتر مرکزی: یزد، صفائیه، بلوار شهیدان اشرف، مرکز فناوری کشاورزی، واحد شماره ۶
کدپستی: ۸۹۱۵۸۱۳۱۵۵

دفتر فروش: یزد، خیابان کاشانی جنب هلال‌احمر، کوچه اخوان پلاک ۱۷، شرکت سیکاس
کدپستی: ۸۹۱۶۷۸۵۳۴۴

info@secaas.ir
۰۹۱۹۷۲۹۶۹۹۱