به نام خداوند یکتا

طرح تولید انباره داده سازمان

و تشریح چارچوب نرم‌افزاری کلان داده

ردهبندی محرمانگی اطلاعات	
■ محرمانه ☐ خیلی محرمانه ☐ سری ☐ فوق سری	
رده‌ی محرمانگی	محرمانه
نویسنده	واحد Big Data، شرکت سیکاس
شماره سند	SCPNKV01.VYB_D980812
مدت اعتبار سند	سه ماه
به هنگام پایان اعتبار، سند: ☐ نابود گردد. ■ بازبینی شود. ☐ در دسترس عموم قرار گیرد. ☐ همیشه معتبر خواهد ماند.	

تهیه و تنظیم:

شرکت سیکاس

(سیناکیان امن‌سازان آریا)

www.secaas.ir

تاریخ انتشار: بهار ۱۳۹۹

این سند دارای دسترسی محدود می‌باشد!

این سند متعلق به شرکت سیناکیان امن‌سازان آریا (سیکاس) بوده و ممکن است حاوی اطلاعات بسیار حساسی باشد. محتویات این سند شامل راهکارهای پیشنهادی سیکاس می‌باشد.

هر نسخه از این سند منحصراً ثبت گردیده است و در صورت نیاز به نسخه‌های اضافی، با شرکت سیکاس به آدرس پست الکترونیک **info@secaas.ir** تماس حاصل فرمایید. هرگونه انتشار و بازتولید غیرمجاز این سند، ممنوع می‌باشد.

با اقدام به مطالعه‌ی مابقی این سند، شما موافقت خود را با شرایط فوق‌الذکر اعلام نموده‌اید.

فهرست مطالب

مقدمه

این مستند با هدف تبیین اهمیت طراحی و تولید یک چارچوب اختصاصی کلان داده در چند بخش تهیه شده است.

در بخش اول مفاهیمی از فرآیند شناخت منابع و اقلام داده ای موجود در سازمان و مراحل شناخت دقیق و شناسنامه‌دار کردن داده‌ها و مسائل(Problems) سازمان آورده شده است که در واقع بیانگر اهمیت شناخت منابع داده‌ای با هدف رفع نقص و پاک‌سازی دستی یا خودکار آن‌ها و تولید انباره داده و Qubeهای اطلاعاتی جهت پاسخ به مسائل سازمان می‌باشد.

در بخش دوم مشخصات چارچوب نرم‌افزاری تهیه شده توسط شرکت سیکاس برای تولید انباره داده و انجام اتوماتیک موارد گفته شده در بخش اول، توضیح داده می‌شود تا خواننده درک صحیحی از آنچه قرار است تولید شود به دست آورد.

در بخش سوم، به تبیین موضوع مهم OpenData و ضرورت پرداختن به آن اختصاص داده شده است. در این بخش بر اهمیت موضوع داده‌های باز و نحوه حرکت به سمت آن در سازمان و تعیین سیاست‌های انتشار آن‌ها پرداخته شده است. موضوعات فنی مورد نیاز این مفهوم در بخش چارچوب نرم‌افزاری لحاظ شده است.

در بخش پایانی نیز به چارچوب اجرایی و فنی پروژه پرداخته شده است.

بخش اول

مفاهیم شناسنامه‌دار کردن، پاک‌سازی، کیفیت داده

و کشف مسائل سازمانی

شماره تماس: ۳۸۳۰ ۱۵۷۴ (۰۳۵)
ایمیل: info@secaas.ir

۱-۱- اهداف طرح

اهداف کلان زیر در صورت انجام برنامه جامع و نقشه راهی که در این پیشنهاد صرفا برنامه آن طراحی می‌گردد مد نظر است:

- تدوین سیاست راهبردی و کوتاه مدت سازمان، متکی بر نتایج حاصل از بکارگیری دانش داده‌کاوی از طریق کشف دانش‌های پنهان درون داده‌ها به جای اتکا بر نگرش‌ها و اطلاعات بخشی و جزیی‌نگر و ناقص.
- طراحی برنامه انجام پیش نیازهای لازم جهت ایجاد سیستم یکپارچه اطلاعات و داده‌های حوزه‌های مختلف شهری و نظام‌مند کردن ورودی و خروجی داده‌ها و اطلاعات.
- امکان رصد کردن لحظه‌ای و به‌روز فرآیندها و اقدامات کل سازمان.
- امکان توسعه دقیق فرآیند تصمیم‌سازی بر مبنای نظام اطلاعات هوشمند به جای اطلاعات ناقص کارشناسان یا مدیران میانی.
- سازگار کردن و به‌روز کردن نظام تصمیم‌سازی سازمان با جدیدترین ابزار و روش پردازش و تحلیل داده و اطلاعات.
- برخوردار نمودن سازمان از دریافت لحظه‌ای دانش‌های پنهان درون داده‌ها از طریق در نظر گرفتن تمامی عوامل و متغیرهای موثر.
- توانمند سازی، سازمان به تحلیل‌های توصیفی، تشریحی، پیش بینی و کنترل به منظورشناخت سریع و پویای وضعیت موجود و یافتن علل و عوامل ایجاد شرایط و پیش بینی آینده و ارائه راهکارهای عملی کنترل و تسخیر آینده.
- به کار گرفتن جدیدترین تجربیات جهانی داده‌کاوی و پردازش اطلاعات در حوزه کاری سازمان

۱-۲- نوآوری طرح

دانش نوین داده‌کاوی (Data mining) دانش در حال توسعه‌ای است که موجب ایجاد انقلاب تکنولوژیک در دهه حاضر گردیده است و بدین‌رو در سال‌های اخیر در دنیا گسترش فوق‌العاده سریعی داشته است. دانش داده‌کاوی فرآیند کشف دانش پنهای درون داده‌ها است که با برخورداری از دامنه وسیع زیرزمینه‌های تخصصی با توصیف، تشریح، پیش بینی و کنترل پدیده‌های گوناگون پیرامونی، امروزه دارای کاربرد بسیار وسیع در حوزه‌های مختلف است.

امروزه داده‌کاوی، مدیران سازمان‌ها را را در موضوعات متنوعی یاری داده بگونه‌ای که استفاده از دانش نوین داده‌کاوی به عنوان جدیدترین و برترین روش حل مسئله برای پایگاه‌های داده سازمان‌ها بسیار حیاتی بوده و بدین جهت امروزه این دانش در تمامی موضوعات و برنامه‌ریزی‌های خرد و کلان، نقش موثر ایفا می‌نماید.

یکی از مهمترین توانمندی‌ها و قابلیت‌های دانش داده‌کاوی نسبت به روش‌های حل مسئله سنتی، امکان پردازش و تحلیل داده‌های کیفی و غیر کمی ا ست که برخی از سازمان‌ها مثل شهرداری‌ها به مانند حوزه مدیریت شهری کشورهای توسعه‌یافته، بدلیل برخورداری از این نوع داده‌ها نیازمند بهره‌گیری از این دانش است.

۱-۳- اهداف و تشریح پروژه

به منظور تحقق اهداف کلان یاد شده ضرورت دارد که برنامه‌ریزی جهت بکارگیری دانش داده‌کاوی در هر سازمانی انجام پذیرد. در گام نخست و فاز صفر این طرح لازم است برنامه جامع و نقشه راهی مطابق با شرایط، ماموریت‌ها و نیازهای سازمان طراحی گردد.

با توجه به گستردگی مجموعه فعالیت‌ها و حوزه‌های مختلف یک سازمان از یک سو و وسعت دامنه کاربردها و توانمندی‌های دانش داده‌کاوی از سوی دیگر، لازم است با بررسی و شناخت حوزه‌های فعالیت تحت پوشش و نیاز هر بخش به دانش داده‌کاوی نسبت به طراحی برنامه جامع و نقشه راه داده‌کاوی در سازمان اقدام گردد.

در این طرح با توجه به ا شراف تیم شرکت سیکاس به توانمندی‌های دانش داده‌کاوی و موارد گسترده به کار گرفته شده در کشورهای تو سعه یافته، برنامه‌های عملیاتی لازم جهت تدوین عناوین پروژه ها و سامانه‌های مورد نیاز جهت بکارگیری دانش داده‌کاوی در بخش‌های مختلف سازمان ارائه می‌گردد که بر ا ساس نقشه راه حا صل از این طرح، سازمان بر نامه کوتاه مدت، میان مدت، بلند مدت خود جهت بهره‌گیری از دانش داده‌کاوی در کلیه حوزه‌ها جهت توصیف، تشریح، پیش بینی و کنترل هوشمند فعالیت‌ها را در دست خواهد داشت.

۱-٤- کیفیت داده

در چرخه عمر هوشمندی کسب و کار، سیستم‌های عملیاتی، نقطه آغازین برای ارائه داده‌ها هستند که در ادامه، مورد تحلیل قرار خواهند گرفت. اگر داده‌هایی که در سیستم‌های عملیاتی ذخیره شده‌اند، به در ستی در

انبار داده‌ها تجمیع نشـوند، امکان تحلیلی صـحیح و جامع از آن‌ها میسـر نخواهد بود. همچنین اگر سـیسـتم‌های عملیاتی دارای خطای داده‌ای باشـند، در زمان نگاشـت (Mapping) و تجمیع (Aggregate) داده‌ها، مشـکلات زیادی بروز خواهند کرد. لذا کیفیت داده‌ها موضوع مهمی در ایجاد انبار داده‌ها و کیفیت تصمیمات حاصله خواهد بود.

چنانچه داده‌ای کاملا اهداف خود را پوشـش دهد، آن‌گاه آن داده‌ها دارای کیفیت شـناخته می‌شـوند. عدم کیفیت داده در سـیسـتم‌های عملیاتی و مکانیزه، یکی از مهمترین عوامل شـکسـت پروژه‌های هوش تجاری (BI) و کلان داده می‌باشـد. چنانچه کاربران احسـاس کنند که داده‌ها و اطلاعات دریافتی آن‌ها از صـحت و دقت کافی برخوردار نیست، اعتماد خود را نسبت به سیستم‌های هوشمندی کسب و کار از دست می‌دهند.

۱-۵- اهمیت کیفیت داده‌ها

به دلایل زیر،کیفیت داده‌ها در تولید انباره داده، از اهمیت بالایی برخوردار است:

- اطمینان مضاعف در تصمیم‌گیری‌ها
- کاهش ریسک تصمیم‌گیری‌ها و در نتیجه کاهش هزینه‌ها
- بهبود در تصمیمات استراتژیک
- پرهیز از تأثیرات مرکب داده‌های آلوده

دلایل ناپاکی و یا بی‌کیفیت‌بودن داده‌ها را می‌توان به صورت زیر عنوان نمود:

- اشکالات سیستمی برنامه‌های عملیاتی و مکانیزه سازمان
- اشکالات در طراحی بانک اطلاعاتی
- استفاده از مقادیر پایه‌ای متفاوت
- وجود شناسه‌های غیر یکتا
- وجود معانی نامفهوم در مقادیر پایه‌ای
- وجود فیلدهای چند منظوره
- تبدیل داده‌ها از سیستم‌های قدیمی و انتقال نادرست اطلاعات از بانک اطلاعاتی قدیمی به جدید
- وجود تعاریف متفاوت یا تعریف نشده یا شفاف نشده از داده‌ها
- گردآوری نادرست داده‌های چند سیستم
- ترکیب نادرست انتظارات کاربران

- داده‌های ورودی ناقص
- فقدان سیاست‌های مدیریت و کنترل صحت داده‌ها ورودی
- وجود چند کپی از داده‌ها در دست چندین نفر و به روزرسانی شخصی آن داده‌ها توسط هر فرد
- ورود سهوی داده‌های اشتباه
- ورود عمدی داده‌های اشتباه (مثلاً کلاهبرداری)

۱-۶- کنترل کیفیت داده‌ها

مالکیت داده یا Data Ownership به عهده واحدهای متولی داده در سازمان است. واحدهای مختلف سازمان در رابطه با داده‌های خود، سطوح دسترسی، نحوه توزیع و قوانین مرتبط سازمانی را بهتر از دیگران می‌دانند.

نظارت بر داده یا Data Stewardship به عهده مدیران واحدهای متولی داده یا واحدهای نظارتی یا ستادی سازمان است. در این سطح از اختیارات، امکان ارائه پیشنهاداتی در خصوص دسترسی، امنیت، توزیع و نگهداشت ارائه می‌شود.

نگهداشت داده‌ها یا Data Custodianship به عهده مدیریت زیرساخت فناوری اطلاعات سازمان است. از آن جایی که این مدیریت صلاحیت لازم در خصوص مهارت‌های فنی لازم، شناخت مفاهیم و رویه‌های امنیت داده، رویه‌های بازیابی داده‌ها و ... را دارا می‌باشد، مسئولیت ذخیره‌سازی، در دسترس قراردادن، پشتیبان‌گیری، آرشیو داده و مواردی این‌گونه را در اختیار دارد.

۱-۷- شناسایی و مستندسازی فراداده‌ها (Meta Data)

- فراداده کسب و کار (Business Metadata): به توصیف معانی هوشمندی کسب و کار و داده‌های انبارشده از دید سازمان می‌پردازد، داده‌هایی چون تعریف داده و سنجه (metric)ها، قوانین کسب و کار، مدل‌های داده‌ای، تعریف مالک/ناظر داده و غیره.
- فراداده فرآیندی (Process Metadata): به توصیف مکان، زمان و چگونگی مراحل دریافت، تبدیل و بارگذاری داده می‌پردازد، داده‌هایی چون نقشه‌های منبع/مقصد، قوانین تبدیل داده، قوانین پاک‌سازی داده و غیره.
- فراداده کاربرد (Application Metadata): به توصیف چگونگی دسترسی و استفاده از داده می‌پردازد، داده‌هایی چون تاریخچه دسترسی به داده، فرکانس، زمان و چگونگی دسترسی به داده و غیره.

- فرادادە فنی (Technical Metadata): به توصیف مکان‌های فیزیکی داده، قالب‌های داده، حجم داده، نام‌های فنی، نوع داده، ساختارهای داده و غیره می‌پردازد.

۱-۸- مدیریت داده‌های اصلی یا MDM (Master Data Management)

در هر سازمان، داده‌های اصلی زیادی وجود دارد که باید برای انتقال در انبار داده‌ها، یکپارچه و Unique شوند. این داده‌ها و اطلاعات ا صلی می‌توانند انواع محصولات، انواع خدمات، تقسیمات جغرافیایی، سرفصل‌های بودجه‌ای، اسامی پروژه‌ها و .... در پایگاه‌های مختلف سازمان باشند که باید یکپارچه‌سازی شوند. به عملیات یکپارچه‌سازی این داده‌های اصلی MDM یا Master Data Management گفته می‌شود.

نمونه رایج و همیشگی مشکلات MDM را می‌توان به صورت زیر دسته‌بندی نمود:

- مدیریت صفات و ساختار فرزند ساختاری بین صفات مرتبط و همچنین یکپارچه‌سازی مقادیر و کدهای مرتبط با صفات در انبار داده‌ها
- مدیریت اطلاعات مرتبط به داده‌های اصلی در دو پایگاه اطلاعاتی مختلف
- مدیریت فرزند ساختارهای غیر یکسان در دو پایگاه اطلاعاتی مختلف
- و مواردی دیگر

البته مشکلات وجود داده‌های اصلی متناقض و یا تکراری به ورود اطلاعات های اشتباهی و همچنین خطاهای سیستمی نیز مرتبط می‌باشد.

۱-۹- پیشنهاد فنی انجام پروژه

سیکاس، فازبندی کلی زیر را به عنوان ساختار تفکیکی اجرای این طرح در نظر دارد. بدیهی است این شکست کار، بسته به نوع فعالیت و تعدد بانک‌های اطلاعاتی هر سازمان، متغیر می‌باشد:

فاز	کد فعالیت	عنوان فعالیت	مستندات قابل ارائه
۱	۱-۱	انجام نشست و مصاحبه با مدیران ارشد سازمان جهت شناخت نیازها و مسایل داده محور حوزه های مختلف سازمان	صورت جلسه
	۱-۲	احصاء نیازها و مسایل داده محور حوزه های مختلف سازمان	فرم شناخت و تحلیل سیستم‌ها
	۱-۳	تصویب نهایی نیازها و مسایل داده محور حوزه های مختلف با تایید مدیران ارشد هر حوزه و تایید نهایی ناظر پروژه	صورت جلسه
	۱-۴	تدوین پیش نیازهای اجرایی پروژه ها	صورت جلسه

فاز	کد فعالیت	عنوان فعالیت	مستندات قابل ارائه
۲	۲-۱	بررسی سیستم‌ها و فرآیندهای داده محور سازمان	فرم شناخت و تحلیل سیستم‌ها
	۲-۲	تهیه شناسنامه نحوه و میزان مالکیت داده های شناسایی شده	فرم مشخصات پایگاه داده
	۲-۳	شناسایی مکعب‌های داده های مبتنی بر هر یک از منابع شناسایی شده	فرم مشخصات کیوب‌های اطلاعاتی
	۲-۴	شناسایی نواقص داده های شناسایی شده	فرم گزارش خطا در داده
	۲-۵	بررسی ابزارهای موجود جمع‌آوری و تجمیع داده‌ای	فرم تشخیص ابزار
	۲-۶	انتخاب ابزار مناسب برای جمع‌آوری داده‌ها	فرم تشخیص ابزار
۳	۳-۱	شناسایی لیست بانک‌های داده ای موجود در سازمان	فرم مشخصات پایگاه داده
	۳-۲	استخراج جدول خصوصیات منابع داده ای سیستم‌های اطلاعاتی سازمان	• فرم مشخصات پایگاه داده • فرم مشخصات KPI
	۳-۳	استخراج جدول خصوصیات منابع داده ای غیر ساخت یافته	• فرم مشخصات پایگاه داده • فرم مشخصات KPI
	۳-۴	استخراج روابط بین داده ای	• فرم روابط بین داده‌ای
	۳-۵	ترسیم شماتیک داده های شناسایی شده و جریان داده ای	• فرم روابط بین داده‌ای
	۳-۶	تهیه دستورالعمل واکشی داده از منابع موجود	• فرم روابط بین داده‌ای
	۳-۷	نصب و استقرار نرم‌افزار جمع‌آوری داده‌ها	• فرم تشخیص ابزار
	۳-۸	اجرای فاز پاکسازی داده‌های جمع‌آوری شده و تغییر شکل داده‌ها	• فرم مشخصات پایگاه داده
	۳-۹	پیکربندی دائمی سامانه‌ها و ابزارهای جمع‌آوری برای بروزرسانی انباره داده	فرم گزارشات انجام کار
	۳-۱۰	نصب و استقرار ابزار مانیتورینگ جریان داده‌ها	فرم گزارشات انجام کار
	۳-۱۱	امکان افزودن منابع داده‌ای بیرونی به انبارداده	فرم گزارشات انجام کار
	۳-۱۲	امکان افزودن رویه‌هایی برای دریافت اطلاعات از منابع اطلاعاتی جدید و افزودن سیاست به روزآوری انباره‌دادها	فرم گزارشات انجام کار
۴	۴-۱	تهیه دستورالعمل برای تکمیل داده های ناقص	فرم گزارش خطا در داده
	۴-۲	شناسایی و پیشنهاد راهکارهایی برای استخراج ساده تر منابع شناسایی شده	فرم مشخصات ETL
	۴-۳	ارائه راهکارهایی برای پاکسازی و رفع نواقص داده های شناسایی شده	• فرم گزارش خطا در داده فرم گزارش خطا در داده
	۴-۴	تهیه شناسنامه کلیه منابع داده ای سازمان با تمامی جزییات و ارزش فیلدها و شناسایی فراداده های مربوطه	فرم مشخصات پایگاه داده
	۴-۵	طراحی عناوین و مشخصات کامل پروژه های داده کاوی مورد نیاز، اثر بخش و قابل انجام در سازمان	فرم گزارشات انجام کار
	۴-۶	تهیه گزارش و عیب‌یابی خطاهای احتمالی در واکشی اطلاعات	فرم گزارش خطا در داده
	۴-۷	تهیه طرح استقرار و ذخیره‌سازی داده‌های سیستم	• فرم مشخصات ETL فرم گزارش کار

شماره تماس: ۳۸۳۰۱۵۷۴ (۰۳۵)
ایمیل: info@secaas.ir

۱-۱۰- مراحل و فازهای اجرای پروژه

بر اساس جدول فوق، فازبندی اجرایی این طرح و اقدامات هر فاز، به شرح زیر می‌باشد:

فاز یک: نیاز سنجی

- بررسی فرآیندهای داده محور سازمان
- انجام نشست و مصاحبه با مدیران ارشد سازمان جهت شناخت نیازها و مسائل داده محور حوزه‌های مختلف سازمان
- احصاء نیازها و مسائل داده محور حوزه‌های مختلف سازمان
- تصویب نهایی نیازها و مسائل داده محور حوزه‌های مختلف سازمان با تایید مدیران ارشد هر حوزه تأیید نهایی ناظر پروژه
- بررسی سامانه‌ها و پژوهش‌های انجام یافته داده محور جهت پرهیز از دوباره‌کاری و اطمینان از نیاز به پروژه‌ها و سامانه‌های مبتی بر داده‌کاوی

فاز دو: امکان سنجی

- مطالعات جامع کاربردهای دانش داده‌کاوی در حوزه‌های کاری سازمان
- ارائه تجربیات عملی و موفق کاربردهای دانش داده‌کاوی در حوزه‌های کاری سازمان بر اساس نمونه‌های موفق انجام شده در سایر سازمان‌ها یا کشورهای توسعه یافته با هدف اطمینان از نتیجه بخش بودن هر کدام از پروژه‌های داده‌کاوی
- تعیین دقیق نیازهای داده‌کاوی اعم از پروژه‌ها و سامانه‌های مبتنی بر داده‌کاوی در حوزه‌های کاری سازمان
- امکان سنجی اجرای پروژه‌های داده‌کاوی شناسایی شده در سازمان با شرط تولید ارزش

فاز سه: بررسی پایگاه داده‌ها و تهیه دیکشنری داده‌ها

- تهیه شناسنامه نحوه و میزان مالکیت داده‌های شناسایی شده
- بررسی و شناخت کلیه منابع داده‌ای، پایگاه داده‌ها و سامانه‌های اطلاعاتی سازمان اعم از مالی، اداری، حمل و نقل، تخصصی حوزه‌های مختلف و ...

شماره تماس: ۳۸۳۰ ۱۵۷۴ (۰۳۵)
ایمیل: info@secaas.ir

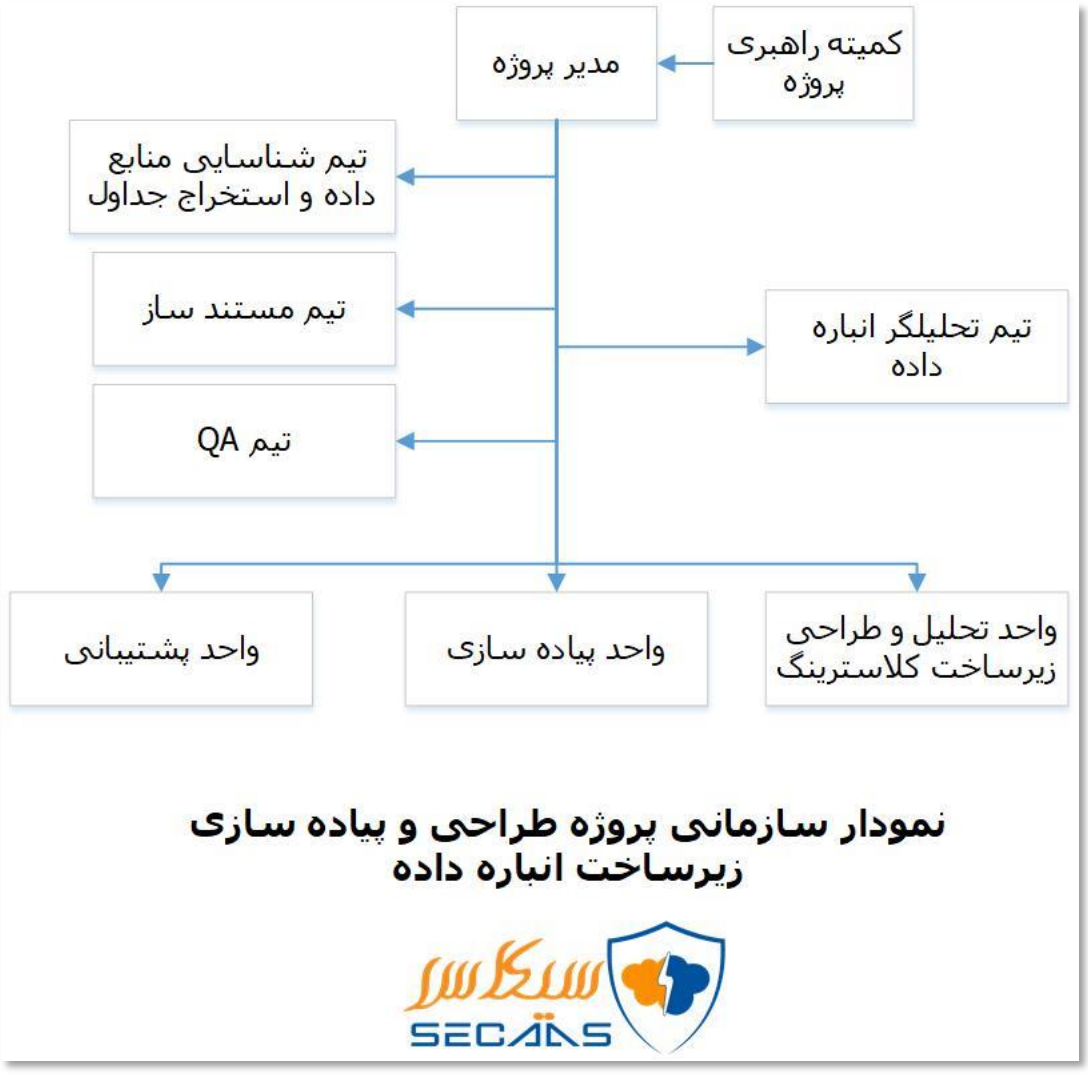
- تهیه دیکشنری داده‌ها برای کلیه پایگاه داده‌ها و سامانه‌های سازمان و روابط بین داده‌ها
- شناسایی مکعب‌های داده‌ای مبتنی بر هر یک از منابع شناسایی شده
- اطمینان از درخور بودن و پشتیبانی داده‌ها از نیازهای داده‌کاوی به نحوی که در صورت نیاز به پاکسازی و رفع نقص داده‌ها، راهکار مناسب به ازاء همه موارد پیشنهاد گردد
- تهیه شناسنامه کلیه منابع داده‌ای به سازمان با تمامی جزئیات و ارزش فیلدها و شناسایی و اتصال فراداده‌های فنی به این منابع

فاز چهار: طراحی، نتیجه گیری و ارائه گزارش نهایی

- طراحی عناوین و مشخصات کامل پروژه‌های داده‌کاوی مورد نیاز، اثر بخش و قابل انجام در سازمان
- تعیین مشخصات فنی پروژه‌ها به تفکیک در قالب RFPهای قابل واگذاری به بخش خصوصی
- تعیین دقیق دستاوردهای مورد انتظار پروژه‌ها به تفکیک
- پیشنهاد دستورالعمل‌هایی به منظور ایجاد قابلیت به اشتراک گذاری منابع داده‌ای
- تدوین پیش نیازهای اجرائی پروژه‌ها
- طراحی جدول زمان‌بندی اجرای پروژه‌ها
- تخمین هزینه پروژه‌های شناسایی شده به صورت نفر ساعت کارشناسی

۱-۱۱- سازمان پیشنهادی انجام پروژه

سازمان پیشنهادی این شرکت جهت انجام این طرح به شکل زیر می‌باشد. بدیهی است بر اساس تعریف پروژه و نیاز سازمان، این ساختار ممکن است تغییراتی داشته باشد.



شکل شماره ۱: نمودار سازمانی اجرای پروژه تولید انباره داده

در ادامه، بخشی از شرح وظایف واحدهای مربوط به سازمان پروژه به اختصار ارائه می شود:

- **مدیر پروژه :**

مدیر پروژه بالاترین مقام اجرایی گروه مجری پروژه و نماینده رسمی شرکت در قرارداد بوده و وظایف و اختیارات زیر را دارا است:

- وظایف مدیریت عمومی پروژه
- برنامه ریزی، هدایت و هماهنگی فرآیندهای فنی و مدیریتی پروژه
- برگزاری، اداره و ارائه گزارش به جلسات کمیته راهبری پروژه، به عنوان نماینده رسمی شرکت
- برگزاری و اداره جلسات کمیته فنی
- تامین و تخصیص منابع لازم برای اجرای پروژه
- گزارش دهی مستمر از پیشرفت کار و مشکلات اجرایی پروژه به هیئت مدیره شرکت
- انتخاب و به کار گماردن اعضای تیم های اجرایی پروژه

- **کمیته راهبری پروژه**

کمیته راهبری، نهاد واسط سازمان اجرایی پروژه کارفرما می باشد و از مدیر پروژه، نماینده رسمی کارفرما و سایر نمایندگان کارفرما و شرکت تشکیل می شود، جلسات این کمیته از زمان آغاز عملیات اجرایی پروژه تا پایان قرارداد به صورت ادواری حداقل هر دو هفته یکبار با مسئولیت مدیر پروژه تشکیل می شود.

وظایف و اختیارات این کمیته عبارتند از:

- ایجاد هماهنگی اجرایی و فنی بین سازمان اجرایی پروژه و عناصر سازمان کارفرما در مراحل مختلف پروژه
- تامین منابع سازمانی لازم در محیط کارفرما جهت پیشبرد پروژه
- نظارت عالیه بر پیشرفت کار مراحل پروژه و بررسی گزارش‌های ادواری

- **تیم تحلیل و طراحی زیرساخت کلاسترینگ**

- انجام مراحل شناخت، تحلیل، طراحی کلی و طراحی تفضیلی
- تهیه مستندات سیستم
- همکاری با تیم مستند سازی جهت تهیه مستندات کاربر
- همکاری با تیم **QA** جهت تهیه طرح‌های آزمون
- مشارکت در مرحله انتقال سیستم

- **تیم پیاده سازی**
  - تهیه نمونه (**prototype**) در مرحله طراحی کلی
  - ساخت پایگاه اطلاعاتی و انباره داده در مرحله طراحی تفضیلی
  - انجام مرحله ساخت
  - اجرای آزمون واحد
  - مشارکت در اجرای آزمون سیستم و انتقال

- **تیم آزمون QA**
  - تهیه طرحهای آزمون با مشارکت تیم تحلیل و طراحی
  - اجرای آزمون یکپارچگی
  - اجرای مرحله آزمون سیستم
  - تهیه و ورود اطلاعات آزمایشی و یا پایه برای انباره داده

- **تیم مستندسازی**
  - تهیه مستندات کاربری پروژه
  - نگهداری و به روز رسانی مستندات خروجی از تیم تحلیل و پیاده سازی
  - همکاری در تهیه مستندات فنی با تیم های تحلیل
  - همکاری با تیم **QA** جهت تهیه مستندات تست سیستم

بخش دوم

مشخصات فنی نرم‌افزار Data Lake

شماره تماس: ۳۸۳۰ ۱۵۷۴ (۰۳۵)
ایمیل: info@secaas.ir

۲-۱- معرفی محصول و قابلیت‌های آن

یکی از نیازهای ضروی سازمان‌ها و نهادهای 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 پرداز شی مورد نیاز برای تأمین نیازهای دانشی سازمان می‌باشد. تنظیم سازوکاری برای نگهداری داده ها بر اساس سری زمانی، تعیین سطح دسترسی افراد به منابع داده‌ای، چگونگی و زمان‌بندی واکشی داده‌ها، شناسایی دقیق محل بروز خطا در زمان انتقال، دریافت شناخت دقیق از کیفیت داده به منظور انتخاب مدل‌های مناسب یادگیری ماشینی و ده‌ها نیازمندی دیگر می‌تواند سازمان را به سمت ایجاد یک بستر اختصاصی و چارچوب مناسب برای بر عهده گرفتن این وظایف حرکت دهد.

نرم‌افزار تولیدی شرکت سیکاس، به منظور پاسخ‌دهی به این مسائل و ایجاد یک بستر مناسب به منظور استاندارد سازی تمامی فعالیت‌های حوزه کلان داده در سازمان، توسط تیم برنامه‌نویسی و کلان داده این شرکت طراحی و تولید شده است. در این نرم‌افزار سعی شده است که عمده دسترسی‌های کاربران تخصصی به داده‌ها در سطح Web Interface تامین گردد که قابلیت‌های مدیریتی و کنترلی مناسبی برای کاربران لحاظ گردد. از موارد شاخصی که در این نرم‌افزار پیش‌بینی شده است می‌توان به قابلیت‌های کلی زیر اشاره کرد:

- ایجاد یک پایپ لاین از مرحله استخراج داده تا لحظه بصری‌سازی
- امکان قابلیت مانیتورینگ داده در تمام دوره حیات آن (LineAge)
- امکان تعیین سطح دسترسی کاربران تا حد فیلد و رکورد به نحوی که کاربران حتی امکان مشاهده کل داده‌ها را نداشته باشند و صرفا با عملیات Sampling مدل‌های خود را اجرا نمایند
- امکان ایجاد کلاسترهای مجازی برای مدیریت بار کاری کاربران حوزه‌های مختلف سازمان برای جلوگیری از تداخل مصرف منابع سخت‌افزاری
- امکان در اختیار داشتن تمامی مدل‌های رایج یادگیری ماشینی برای جلوگیری از کدنویسی مجدد
- امکان استفاده مجدد کاربران از نتایج کارهای قبلی و سایرین
- 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

این نرم افزار دارای اجزا و قسمتهای زیر می باشد:

۲-۱-۱- بخش استخراج داده ها

- 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 داده های 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 استخراج‌شده 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 پیشنهادهای پاک‌سازی روی داده‌ها و فیلدها از قبیل عدم رعایت هم‌سانی واحدزمانی، عدم رعایت هم‌سانی واحد وزن، عدم رعایت واحد طولی، عدم رعایت واحدپولی، عدم رعایت هم‌سانی سایر پارامترها، قرار نداشتن مقادیر در محدوده مجاز، عدم رعایت یکتایی مقادیر و ..
- به ازای هر یک از ستون‌ها، اطلاعات آماری زیر مرتبط با نوع آن فیلد استخراج می‌شود:
  - Essentials: type, unique values, missing values
  - Quantile statistics like minimum value, Q1, median, Q3, maximum, range, interquartile range
  - Descriptive statistics: like mean, mode, standard deviation, sum, median absolute deviation, coefficient of variation, kurtosis, skewness
  - Most frequent values

- Histogram
- Record Count: The total number of records in the column
- Unique Values: The distinct count of unique values in a column
- Empty Strings: The count of empty strings in a column
- Null Values: The count of null values in a column
- Percent Fill: The percentage of records that are not empty string or null
- Percent Numeric: The percentage of records that are numeric
- Max Len: The max length of data

۲-۱-۲- بخش ساخت ELT و CUBE های داده ای

- شناسایی بعد داده‌ها از نظر ویژگی، تعداد مقادیر، مجموعه داده‌ها
- امکان ایجاد کپی از جداول و Data Frame های استخراج‌شده
- 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 ...
- استخراج روابط بین جداول و نمایش گرافیکی آن‌ها
- امکان مشاهده گرافیکی میزان نقص داده‌های جداول مبتنی بر نقص رکورد یا فیلد (به قابلیت Drill-Drown) 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 را رد **UC** ها یا DF های شناسایی‌شده
- امکان مشاهده میزان نقص داده DF های هر UC به صورت گرافیکی

۲-۱-۳- بخش تحلیل و آنالیز های ML داده ها

- امکان مشاهده Sample داده از DF ها یا Cube ها
- امکان اجرای الگوریتم‌های کاهش ابعادی هم‌چون PCA و SVD روی DF ها یا Cube ها و ذخیره DF یا Cube جدید
- 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
- امکان کار با انواع داده‌های GIS و جغرافیایی
- امکان کار با Meta Data یا آمارهای کلان می‌توان مقادیر زیر را محاسبه کرد:
  - **بسامد 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 وب‌سرویس و API بر روی موجودیت‌های سیستم اعم از DF و جداول، مکعب‌های داده‌ای و ...
- امکان ساخت Df های جدید از DF های استخراج شده در قالب یک Cube یا Data Mart
- اتصال موتور کدنویسی به موتور Machine Learning اسپارک برای ایجاد قابلیت استفاده برنامه‌نویس در محیط ادیتور از الگوریتم‌های مربوطه
- امکان کدنویسی برای استخراج منابع داده‌ای از شبکه‌های اجتماعی بر اساس کلمه کلیدی
- امکان ذخیره‌سازی نتیجه یک کد نوشته شده در قالب DF به نحوی که قابل استفاده برای سایرین باشد.

۲-۱-٤- 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** Chernoff Faces

امکان بصری‌سازی نتایج مدل‌های یادگیری ماشینی تولید شده

۲-۱-۵- بخش های 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 اتصال سیستم به یک زمان‌بند مانند Apache Nifi

امکان تعریف زمانبندی اتصالات به منابع داده‌ای و بازه‌های زمانی به روزرسانی دریاچه داده

امکان مشاهده گزارش نتیجه واکشی داده‌ها از نظر سلامت اجرا یا خطا در زمان واکشی اطلاعات

قابلیت‌های پیش‌بینی‌شده برای ارائه خدمات به شرکت‌های استارتاپی به صورت زیر پیش‌بینی شده است:

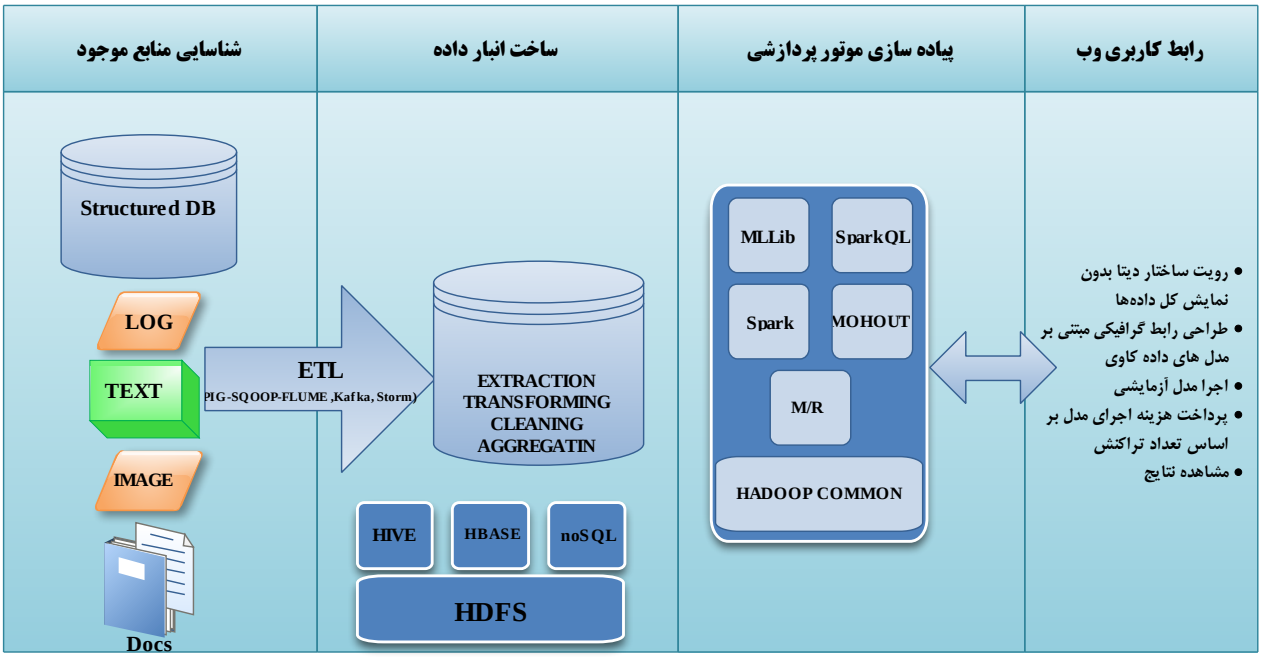
امکان ساخت انواع گزارش‌ها از داده‌ها و اطلاعات استخراج شده از مدل طراحی‌شده

امکان ذخیره‌سازی مدل‌ها

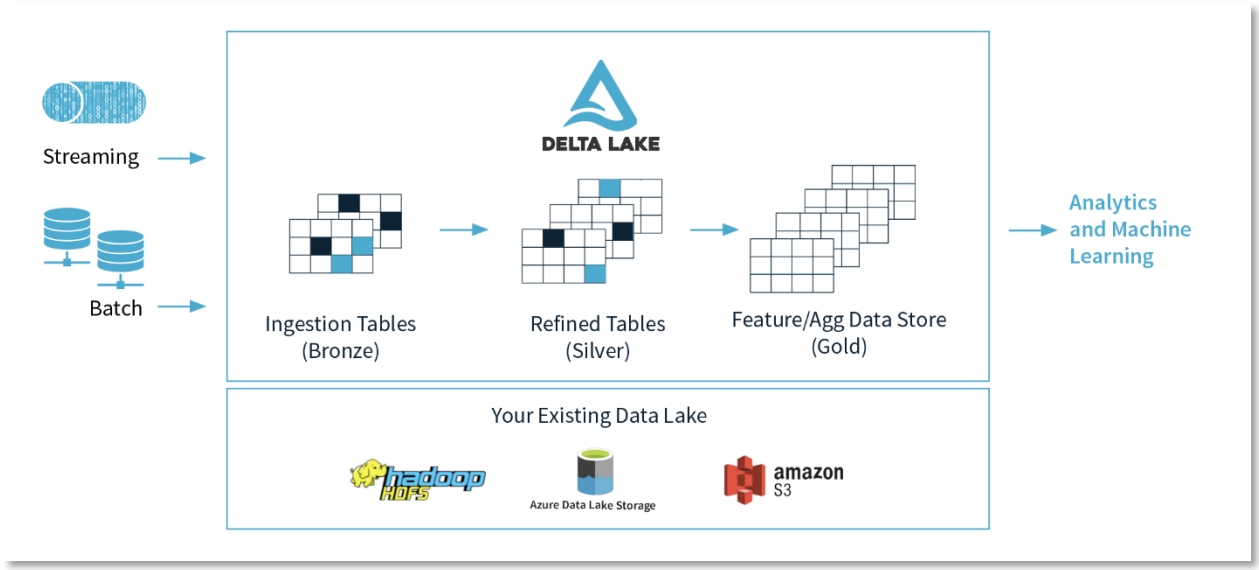
امکان به اشتراک‌گذاری یک مدل با سایر کاربران یا گروه‌های کاربری

- امکان ثبت‌نام کاربران انسانی جدید و تایید عضویت و سطح دسترسی توسط مدیر یا مدیران سامانه
- امکان بلاک‌کردن یک یا گروهی از کاربران
- امکان تعریف Application هایی که می‌توانند از خروجی مدل‌ها، داده دریافت کنند و فرایند احراز هویت آنان و انتساب آن App به یک کاربر خاص
- امکان تعریف سیاست‌های دریافت اطلاعات توسط Application های مذکور (بازه‌های زمانی اتصال، تعداد دفعات دسترسی در یک بازه زمانی، تعداد رکوردهایی که در هر بار درخواست می‌توان پاسخ داد و ...)
- امکان تعریف سیاست‌های دریافت هزینه برای اجرای تراکنش‌های یک یا گروهی از کاربران (بر اساس تراکنش، حجم داده، زمان اجرا و ...)
- امکان شارژ حساب کاربری از طریق کارت‌های شتاب (توسط کاربر) یا به صورت اعتباری (توسط مدیر)
- کسر خودکار از حساب فرد پس از اجرای Query ها (Per record or transaction)
- امکان اخذ گزارش‌های مدیریتی مختلف از قبیل مشاهده کاربران فعال، کاربران در حال اجرای Query، موجودی فعلی کاربران، درآمد حاصله، ریز کسر از حساب، حجم و تعداد تراکنش‌های اجرا شده توسط کاربر(ان)، ریز فعالیت یک کاربر، حجم و نوع داده‌های موجود در انباره داده، حجم داده‌های ارسالی و دریافتی به تفکیک زمان و IP و ...
- امکان مشاهده لیست و ساختار داده‌های (Data Sets) موجود در دریاچه داده به صورت محدود
- امکان تعریف داده‌هایی که می‌تواند توسط کاربران یا گروه خاصی از کاربران دیده شود و سطح دسترسی به آن داده (دانلودشدن، مشاهده ساختار، تعداد رکورد قابل مشاهده، حداکثر حجم قابل ارسال در هر بار پاسخ به Query، حداکثر حجم قابل ارسال در هر رکورد و ...)
- امکان ساخت خودکار یک وب سرویس مبتنی بر مدل ساخته شده و منطبق بر سیاست‌های تعیین‌شده توسط کاربرانی که امکان دریافت داده از طریق Application را دارند (احراز هویت این وب سرویس‌ها بر اساس اطلاعات کاربری مالکان App مربوطه)

۲-۲- معماری مفهومی سیستم

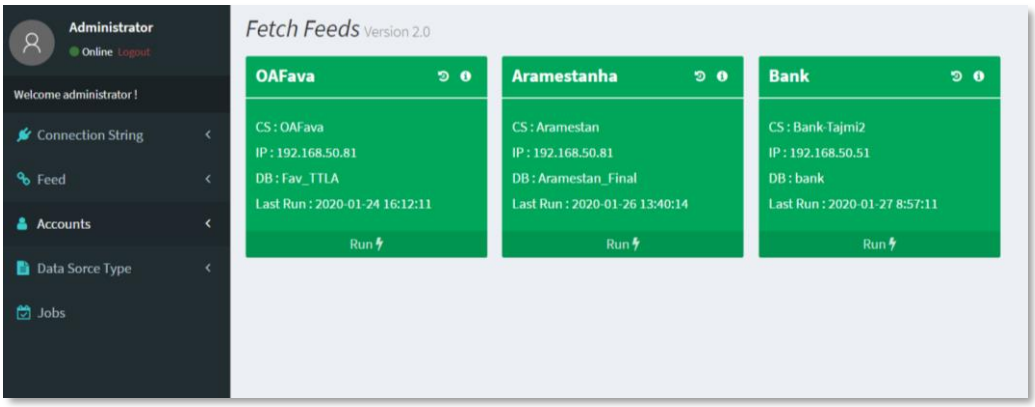


شکل شماره ۲: معماری مفهومی سیستم

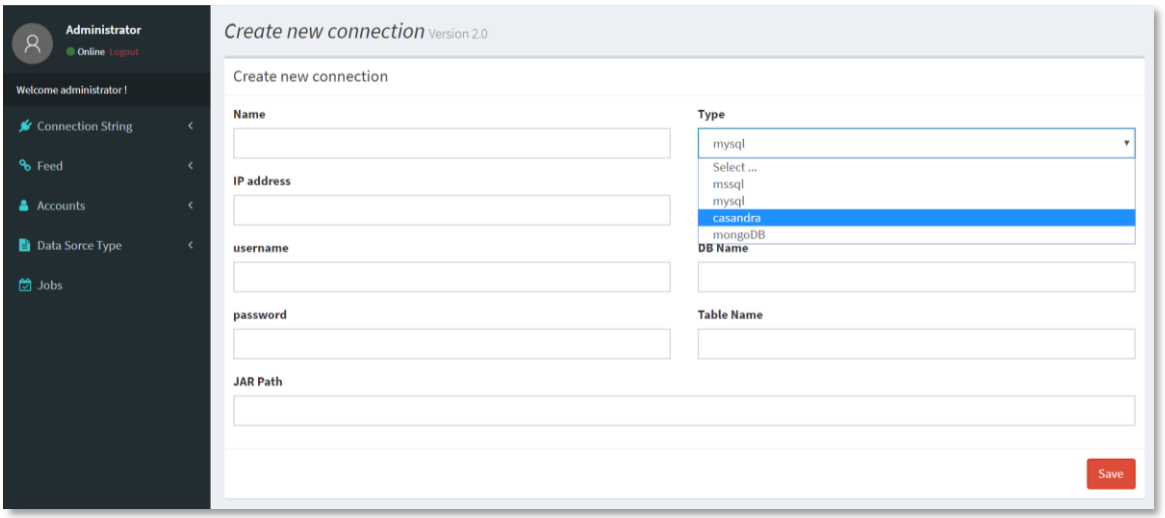


شکل شماره ۳: شمایی از عملیات بر روی داده‌ها

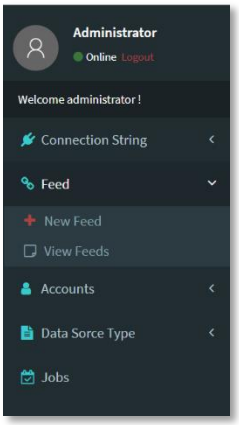
۲-۳- بخش‌هایی از رابط کاربری نرم‌افزار

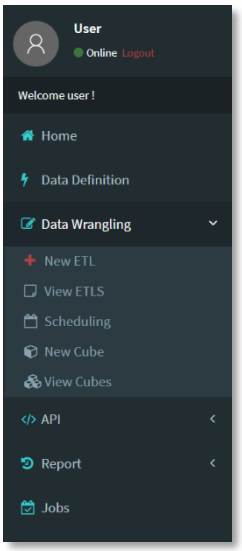


شکل شماره ۴: وضعیت دیتابیس‌های فید شده جهت استخراج متادیتا

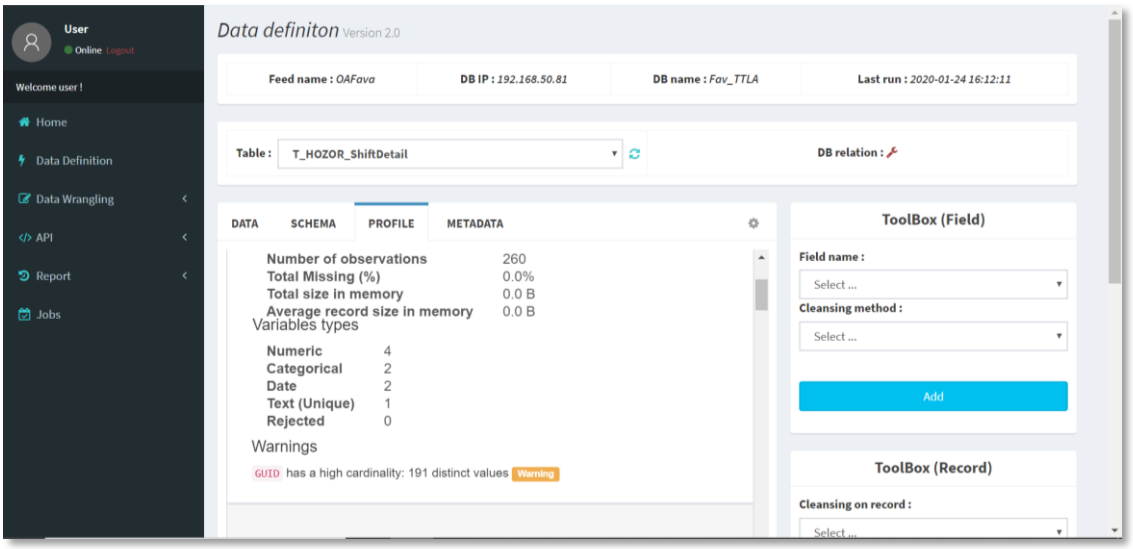


شکل شماره ۵: تعریف اتصال جدید به یک دیتابیس جهت انجام عملیات Profiling

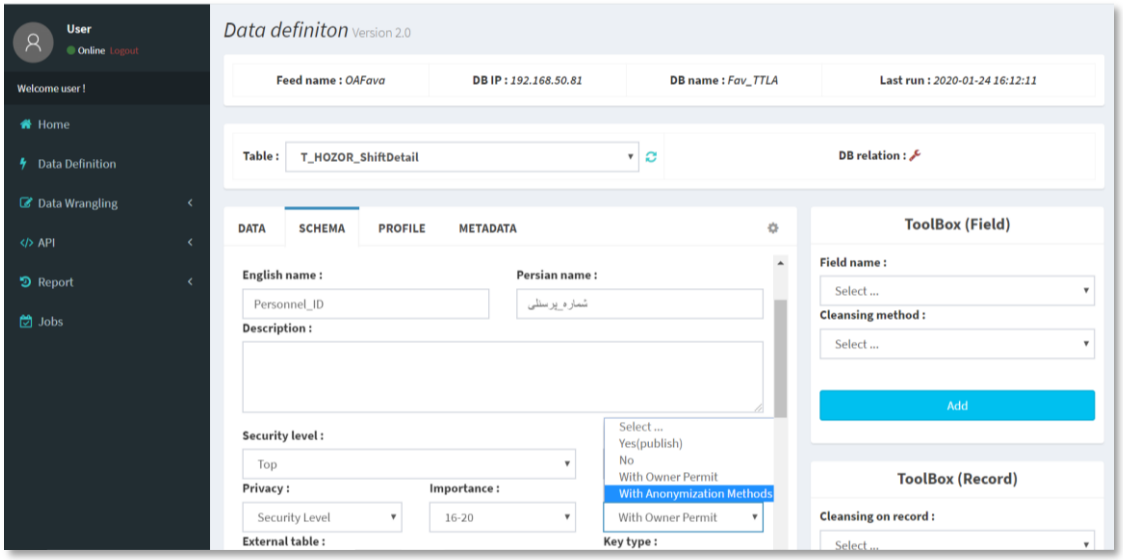




شکل شماره ۶: منوهای دو حالت کاربری User و Administrator



شکل شماره ۷: خروجی گزارش Profiling دیتابیس



شکل شماره ۸: فرم شناسنه‌دار کردن جدول اطلاعاتی (**Data Definition**)

بخش سوم

داده‌های باز و سیاست انتشار داده‌ها

شماره تماس: ۳۸۳۰ ۱۵۷۴ (۰۳۵)
ایمیل: info@secaas.ir

"داده باز" مفهوم بسیار جدیدی است. ریشه این مفهوم در اعتقاد به این امر است که شهروندان باید به حجم انبوه اطلاعاتی که به طور مرتب توسط نهادهای دولتی جمع‌آوری می‌شوند، دسترسی داشته باشند. در اواخر دهه نخست قرن بیست و یکم میلادی، دولت‌ها و مؤسسات به تدریج دسترسی افراد به چنین منابعی را گسترش دادند. نخستین قوانین دولتی مربوط به داده‌های باز در سال ۲۰۰۹ شکل گرفت. امروز بیش از ۲۵۰ دولت در سطوح ملی، منطقه‌ای و شهری، حدود ۵۰ کشور توسعه‌یافته و در حال توسعه، و سازمان‌هایی چون بانک جهانی و سازمان ملل متحد، برنامه‌هایی برای پیاده‌سازی الزامات داده‌های باز راه‌اندازی نموده‌اند و هر سال نیز تعداد بیشتری از این برنامه‌ها آغاز به کار می‌کنند.

داده‌ها در صورتی «باز» محسوب می‌شوند که هر شخصی بتواند به صورت رایگان، بدون محدودیت و برای هر منظوری از آن‌ها استفاده کند و آن‌ها را مجددا توزیع نماید. اگر چه حجم وسیعی از داده‌ها در وبگاه‌های دولتی منتشر می‌شوند، اما اکثر داده‌های منتشرشده صرفاً به صورت مستنداتی برای مطالعه در اختیار کاربران قرار می‌گیرد و امکان استفاده از آن‌ها برای اهداف دیگر وجود ندارد. داده برای آنکه «باز» به حساب آید، باید به صورتی منتشر شود که امکان دانلود در قالب‌های باز را داشته باشد، توسط نرم‌افزار قابل خواندن باشد و کاربران دارای این حق قانونی باشند که از آن بارها استفاده نمایند.

برای شناخت بیشتر این مفهوم نیاز است تا برخی عبارات و مفاهیم تعریف گردند که در ادامه ارائه می‌گردد.

۳-۱- تعریف داده

امروزه همه چیز داده (Data) حساب می‌شود. از شن‌های ریز سواحل گرفته تا منظومه‌ها و سیارات کهکشان‌ها. اما اصولا داده را مواد خامی می‌دانند که یک واقعیت را توصیف می‌کند. این مواد خام متغیرهای کمی(اعداد و..) یا کیفی(حروف و نمادها) هستند. وقتی داده‌ها کنار هم قرار گرفته و معنادار (پردازش) می شوند، در اینجا دیگر داده‌ی خام نیستند، بلکه به اطلاعات (Information) تبدیل شده است.

۳-۲- انواع داده‌ها

داده‌ها می‌توانند انواع گوناگونی داشته باشند از قبیل:

- داده‌های آمار و ارقام
- داده‌های جغرافیایی

- داده‌های حمل و نقل
- داده‌های مالی و بانکی
- داده‌های فرهنگی
- داده‌های علمی
- داده‌های آب و هوایی
- داده‌های طبیعی و محیط زیستی

۳-۳- تعریف فراداده (Meta Data)

هر مطلبی که توضیحی راجع‌به داده‌ها بدهد را فراداده **(Meta Data)** می‌گویند. فراداده‌ها یا در مشخصات فایل داده هستند، یا به‌صورت جداگانه در صفحه دانلود و بخش توضیحات نوشته می‌شوند. فراداده‌ها معمولا شامل: منبع و تهیه‌کننده فایل داده، منتشرکننده داده، تاریخ تهیه و انتشار، تعاریف و مفاهیم، واحد اندازه‌گیری و غیره می‌شود.

۳-٤- تعریف داده بسته (Closed Data)

هر داده‌ای که برای دسترسی به آن محدودیت و هزینه وجود دارد را داده بسته می‌گویند.اگر داده‌ای یک یا تعدادی از ویژگی‌های زیر را داشته باشد، «بسته» حساب می‌شود.

- برای هر بار دسترسی به داده‌ها، مجوز و اجازه نامه لازم است.
- داده‌ها فقط برای اعضای یک گروه یا وب‌سایت نمایش داده می‌شوند. یعنی برای دسترسی به داده یا باید عضو سایت شوید یا هزینه‌ای را پرداخت کنید.
- داده‌ها رمزگذاری یا محدود شده‌اند.
- داده‌ها شامل قانون کپی‌رایت می‌شوند. یعنی در هر بار استفاده از داده‌ها، نام منبع باید ذکر شود.
- موتورهای جست‌وجو، سایت‌ها و API برای دسترسی به داده با محدودیت روبرو هستند.
- از نظر زمان استفاده می‌توانند دارای محدودیت باشند.
- داده‌ها قابل تغییر و ترکیب نیستند.
- داده‌ها در فرمت‌های گوناگونی ارائه نمی‌شوند.

۳-۵- تعریف داده باز (Open Data) و فواید آن

داده باز، داده‌ای است که هر فردی می‌تواند به‌صورت آزاد و رایگان از آن برای هر مقصودی بدون نیاز به مجوز یا اجازه‌نامه استفاده (استفاده مجدد، توزیع) کند.

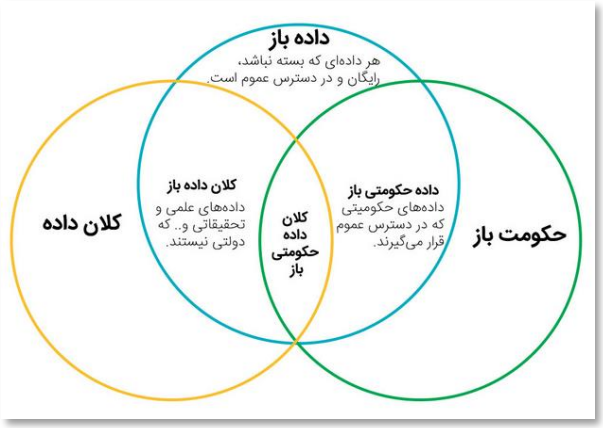
داده باز آن داده‌ای است که امکان استفاده یا باز استفاده به معنی پردازش و ترکیب با دیگر داده‌ها برای دستیابی به داده‌های جدیدتر آزاد و رایگان هرکس از آن برای مقاصد قانونی و مشخص ممکن باشد.

۳-۵-۱- دسترس‌پذیری

اطلاعات، ترجیحا رایگان و در غیر آن با هزینه‌ای معقول، قابل بارگیری از اینترنت و موجود در فرمت‌های قابل پردازش و محاسبه باشد. داده باید ذیل قوانینی منتشر شود که اجازه قانونی استفاده، باز استفاده و انتشار خروجی آن داده شده باشد.

۳-۵-۲- دسترسی همگانی

هرکسی بتواند به این اطلاعات دسترسی داشته باشد. نباید دسترسی به اطلاعات به گروه‌های خاصی محدود شود. در دسترس قراردادن داده‌ها به شکل گسترده و با قابلیت استفاده آسان، منافع قابل توجهی به همراه دارد. داده‌ها می‌توانند موجب تسهیل در ارائه خدمات دولتی، ایجاد فرصت‌های اقتصادی، ترغیب نوآوری، بهبود امنیت عمومی و کاهش فقر شوند. با رشد روزافزون جامعه تأثیرپذیر از داده‌های باز و با کشف فرصت‌های جدید در این زمینه، دولت‌ها و مؤسسات سراسر دنیا علاقه‌مند شده‌اند که برنامه‌های داده باز را آغاز نمایند و یا برنامه‌های موجود را گسترش دهند. فهم کامل پیچیدگی و ظرفیت‌های گسترده داده‌های باز، که از محیط «باز» مجوز نرم‌افزارها منشعب شده است، زمان می‌برد. از آن‌جا که بحث داده‌های باز همچنان در مراحل اوّلیه خود است، بهترین روش ها برای به کارگیری در این حوزه هنوز به طور کامل شناسایی نشده‌اند و جوامع تخصصی پیرامون این موضوع به تازگی در حال شکل‌گیری‌اند.



شکل شماره ۹: جایگاه داده باز

۳-۶- سطوح پنج‌گانه انتشار داده

داده‌های باز برای انتشار سطوح پنج‌گانه‌ای را دارند که در ادامه به معرفی آن پرداخته می‌شود.

کیفیت انتشار داده	عنوان سطح انتشار	نکات + ویژگی‌های اصلی	نمونه قالب‌ها
★	**قالب‌های تصویری**	صرفا قابلیت خوانش انسانی[۵]	PDF, JPEG, PNG
★★	**قالب‌های متنی**	قابلیت کپی‌برداری و جستجو توسط انسان	DOC, HTML
★★★	**قالب‌های مشترک**	قابلیت خوانش و پردازش انسان و نیز ماشین	XLS, CSV, XML
★★★★	**امکان فراخوان ماشینی**	قالب‌هایی با قابلیت پردازش و فراخوانی	API (JSON)[۶]
★★★★★	**رعایت استانداردهای داده متصل**[۷]	تکمیل جورچین دادگان موضوع در ارتباط با دیگر داده‌های موجود	API + Linked Data

جدول شماره ۱۰: سطوح داده‌ها

۳-۶-۱- سطح اول (۱ ستاره)

در سطح اول (یک ستاره) نوع و فرمت داده اهمیت نداشته و مهم این است که داده‌ها با مجوز باز (open license) در دسترس همه قرار بگیرند.

مثلا اگر شـما داده‌هایتان برای همه افراد و کاربران وب منتشـر کنید، این داده‌ها در سطح اول قرار می‌گیرند.

۳-۶-۲- سطح دوم (۲ ستاره)

در سطح دوم، داده هایی قرار می‌گیر ند که دارای ساختار های صفحه‌گسترده و جدولی مثل اکسل (XLS)باشند.

۳-۶-۳- سطح سوم (۳ ستاره)

سطح سوم مشابه سطح دوم است، یعنی داده‌های مرتب‌شده با ساختار جدول، اما تفاوت ا صلی در فرمت داده‌ها ست. به‌عنوان مثال، داده‌های با فرمت XLS را فقط با مایکروسافت اکسل می‌توان مشاهده و ویرایش کرد. فرمت‌هایی مثل CSV عمومی بوده و همه می‌توانند از آن استفاده کنند. پس اگر داده‌های اکسل (XLS) خود را به CSV تبدیل کنید، سه ستاره می‌گیرند. ضمن آن‌که این داده‌ها ماشین‌خوان هم هستند.

۳-۶-۴- سطح چهارم (۴ ستاره)

در سطح چهارم (چهار ستاره)، داده‌ها نه در فایل‌ها، بلکه در صفحات وب منتشر می‌شوند و هر کدام آدرس اختصاصی خود را دارند. یعنی برای مشاهده داده‌ها نیازی به فایل و نرم‌افزار ندارید چرا که داده‌ها از طریق کدهایHTML وXML نوشته می‌شوند. در این حالت می‌توان داده‌ها را به یکدیگر لینک کرد.

۳-۶-۵- سطح پنجم (۵ ستاره)

در سطح پنجم همه‌ی داده‌ها باز بوده و به یکدیگر لینک و مرتبط هستند. در این‌جا شبکه‌ای از داده‌ها را داریم که به یکدیگر وصل هستند.

با این تعریف، اطلاعات غیر دیجیتال، داده استاندارد نیستند. برای استانداردسازی داده‌ها، قالب‌های خاصی تعریف شده است که تبدیل داده‌ها به آن قالب‌ها امکان فعالیت ماشینی و برنامه‌ریزی شده را بر روی داده‌ها فراهم می‌کند.

هر چند حداقل لازم برای انتشار داده‌های باز، سطح دوم یا نیمه ساختاریافته است، اما هرچه داده‌ها ساخت‌یافته‌تر و بسط‌پذیرتر باشند، آن داده امکان استفاده‌های بیشتر و متنوع‌تری خواهد داشت. استانداردهای تبادل داده در تلاش هستند که تمام داده‌ها پس از آن که الزامات سطح چهارم که امکان خوانش و فراخوانی کامل ماشینی است را کسب کردند، در سطح پنجم (اصطلاحا 5 ستاره) منتشر شوند تا چرخه داده باز و فعالیت‌های مرتب بر آن تکمیل گردد.

سطوح اول، دوم و سوم توسط انسان قابل خوانش است، در حالی‌که سطوح سوم، چهارم و پنجم از قابلیت خوانش ماشینی برخوردار است.

در سطح پنجم به علت فراهم‌شدن بستر فنی لازم و همچنین انجام‌شدن داده‌کاوی‌های بسیار، خروجی مورد نیاز ماشین و انسان فراهم است و فرآیند بهره‌برداری از اطلاعات می‌تواند به صورت دوسویه ادامه پیدا کرده و کاوش‌های جدیدی انجام شود.

داده باز متصل، مرحله نهایی فرآیند داده باز است. در این مرحله تمام داده بازهای در ارتباط با یکدیگر قرارگرفته و به تکمیل هم کمک می‌کنند. این توانایی موجب می‌شود اطلاعات گسترده‌ای را بتوان به‌صورت خودکار بین تمامی کامپیوترها، به اشتراک گذاشت. از آن گذشته، باعث می‌شود تا اطلاعات از منابع گوناگون به هم مرتبط شوند و بتوان روی آن‌ها کاوش‌های گوناگونی را اعمال کرد.

۳-۷- ترکیب داده

مرحله بعدی ترکیب داده‌هاست. گاها لازم است خدماتی در وب ارائه شود تا داده‌های مختلفی که (عمدتا به دلیل ضعف در قوانین مادر یا در اجرای آن‌ها) در وب و به صورت پراکنده ارائه شده است، را به صورت یکجا و منسجم گردآوری و ارائه نمایند. کشورهایی که در حوزه داده باز پیشتاز هستند، پرتال واحدی را برای ارائه اطلاعات معرفی نموده‌اند. در غیر این صورت ممکن است برخی از نهادهای مردمی در راستای چنین ضرورتی اقدام کرده باشند.

۳-۸- ارائه تحلیل و یا ابزارهای تحلیل

دسترسی به داده خام بسیار ارزشمند است، اما همچون مواد اولیه و خام آشپزی می‌ماند! در این مرحله نیازمند آشپزهایی حرفه‌ای است که بتوانند با ترکیب داده‌ها، تحلیل‌هایی ارزشمند را ارائه نمایند. در سطح بالاتر بسیار مفید خواهد بود اگر ابزارهایی در وب ارائه گردد که مخاطبان خود بتوانند با ترکیب دلخواه داده‌ها، به تحلیل‌هایی جدید و متناسب با نیاز خود دست یابند. سرویس‌های مختلف وب، چنین تحلیل‌هایی را به صورت رایگان و البته تحلیل‌های پیچیده خود را با نیت تجاری بر روی وب ارائه می‌نمایند.

بخش چهارم

روال اجرایی طرح

شماره تماس: ۳۸۳۰ ۱۵۷۴ (۰۳۵)
ایمیل: info@secaas.ir

این طرح پس از عقد قرارداد اجرا، امضـای قرارداد عدم افشـاء و ابلاغ آن به شـرکت، طی مراحل زیر به انجام خواهد رسید:

- نصب نرم‌افزار بر روی سرور کارفرما و ایجاد دسترسی‌های لازم جهت ارتباط با آن.
- راه‌اندازی یک سرور موقت جهت قرار دادن یک نسخه کپی یا نمونه از بانک‌های اطلاعاتی سازمان بر روی آن، جهت ارتباط نرم‌افزار به بانک‌های اطلاعاتی. (نیاز به این سرور به آن دلیل است که اتصال مستقیم و آنلاین به بانک‌های اطلاعاتی در حال کار، دارای ریسک بوده و از نظر امنیتی، قابل قبول نمی‌باشد.)
- انجام عملیات پروفایلینگ بانک‌های اطلاعاتی توسط نرم‌افزار.
- بررسی بانک‌های اطلاعاتی پروفایل شده و جداول آن‌ها توسط کارشناسان شرکت و پاک‌سازی و شناسنامه دار کردن آن‌ها و ثبت پیشنهادات و گزارشات مربوطه در نرم‌افزار.
- (همزمان با مراحل ۳ و ۴) برگزاری جلسات حضوری با واحدهای مختلف سازمان و کارشناسان مرتبط با بانک‌های اطلاعاتی، جهت کشف مسائل و مشکلات (Problem) سازمان که لازم است از طریق این پروژه، به آنها پاسخ داده شود.
- تهیه و ارائه شناسنامه مسائل کشف شده سازمان
- ساخت کلاسترهای مورد نیاز انباره داده (برای این منظور نیاز به حداقل ۳ سرور سخت‌افزاری با مشخصات لازم برای تجمیع و آنالیز همه داده‌های سازمان می‌باشد)
- تجمیع داده‌های پاک‌سازی شده و ساخت انباره داده
- تعریف مسائل در نرم‌افزار و ساخت مکعب‌های اطلاعاتی (Cube) بر اساس نیاز مسائل و ارائه API جهت استفاده در نرم‌افزارهای Visualization
- آموزش و تحویل پروژه

مدت زمان اجرای این طرح کاملا بســتگی به به حجم داده و تعداد پایگاه داده‌های موجود در ســازمان دارد و ممکن ا ست از ۶ ماه تا یک سال طول بکشد. به همین ترتیب، هزینه اجرای این طرح نیز متغیر و وابسته به حجم داده‌ها دارد. اینکه سازمان صرفا بخواهد تا فاز تولید انباره داده و ساخت کیوب‌های اطلاعاتی پیش برود یا اینکه تمایل به اجرای فاز داده کاوی (Data Mining) و ســایر فازهای مرتبط با این داده‌ها نظیر داده‌های باز (Open data) برود یا نه هم، در زمان و مبلغ اجرا تاثیرگذار خواهد بود.

با تشکر

تیم کلان داده شرکت سیکاس