دانشکده مهندسی کامپیوتر
گروه نرم افزار

سمینار کارشناسی ارشد تحت عنوان :

تحلیل و پردازش داده های تاریک با هدف تبدیل آن به داده های روشن

نگارش:

مهدیه یحیائیان

استاد راهنما:

جناب آقای دکتر شایق

خرداد ۹۷

چکیده

افزایش نرخ تولید داده ها امروزه موجب به وجود آمدن چالشی با عنوان کلان داده ها یا داده های بزرگ شده است. در این گزارش در مورد چالش های داده های تاریک صحبت می شود. داده های تاریک زیر مجموعه داده های بزرگ می باشند و شامل داده های ساخت یافته و داده های بدون ساختار است که البته بخش اعظم آنها را داده های بدون ساختار تشکیل می دهند. منظور از داده های بدن ساختار صرفا داده هایی نیست که در پایگاه داده های رابطه ای جای نمی گیرند، بلکه منظور تمام داده هایی است که بدون توضیح و فرا داده منتشر شده اند. داده-هایی که برای درک آنها نیاز به حضور یا برقراری ارتباط با تولید کننده وجود دارد یا در صورت عدم حضور متخصص و تولید کننده داده رمز گشایی این داده ها زمان بر و پرهزینه است و اکثرا از آن چشم پوشی می شود. در این گزارش به تعریف این داده ها و دلایل توجه به آنها و مطرح کردنشان به عنوان یک چالش، پرداخته شده است. برای حل این چالش دو دیدگاه معرفی شده است که یکی در تلاش است از ایجاد داده های تاریک جلوگیری کند و دیگری درصدد است داده های تاریک را روشن و قابل استفاده کند. نظریه های متعددی در حوزه ی داده های تاریک وجود دارد و از ابعاد مختلف در علوم مختلف مورد بررسی قرار گرفته است اما تقریبا در بیشتر موارد تنها به عنوان یک موضوع و چالش مطرح شده است و تا به امروز روشی که بتوان توسط آن هر نوع از این داده ها را در دسترس ساخت و روشن و قابل استفاده کرد ابداع نشده است.

بحث داده های تاریک بسیار گسترده و نه تنها در حوزه علوم کامپیوتر و نرم افزار بلکه از حوزه های روانشناسی، جامع شناسی و حقوق هم باید بدان پرداخت که البته در این گزارش این موضوع تنها از منظر علم نرم افزار مورد بررسی قرار گرفته است. هر چند که مطالعه ی کار اندیشمندان و متخصصان علوم مختلف در این باره کمک می کند تا بتوان در این حوزه کاری اساسی انجام داد و از نابودی دریای عظیمی از اطلاعات جلوگیری کرد.

کلمات کلیدی: کلان داده ها، داده های تاریک، داده های بدون ساختار، داده کاوی، تحلیل و آنالیز داده های بزرگ

فهرست مطالب

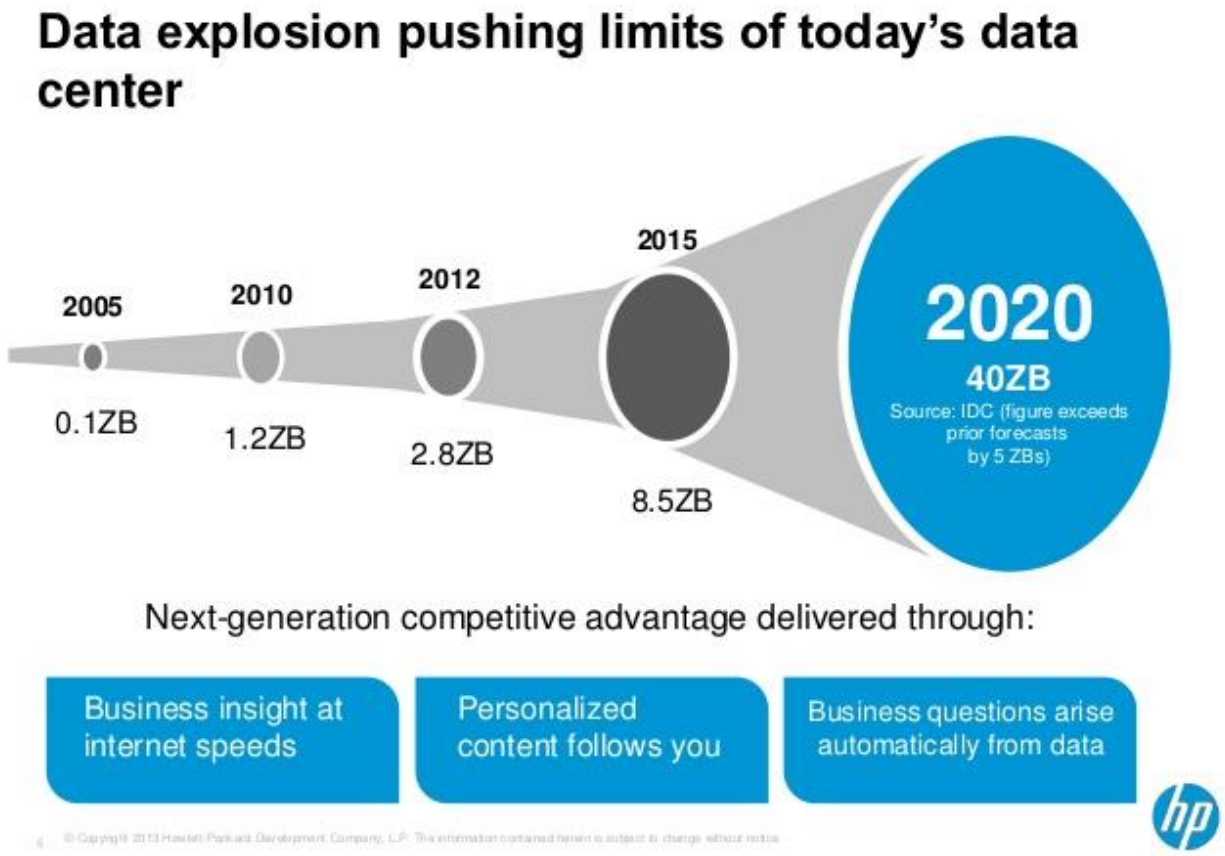
فهرست شکل ها

فصل اول: مقدمه

امروزه همگی با مفهوم داده و اطلاعات آشنایی دارند و استفاده از شبکه های اجتماعی، انواع پیام رسان ها و فروشگاههای اینترنتی امری معمول و متداول است. با متداول شدن استفاده از ابزارهایی به این شکل، روز به روز بر حجم داده های تولید شده از منابع مختلف افزوده می شود و نگهداری از آن سخت و پردازش آن سخت تر می شود. از این رو تمهیداتی اندیشیده شده است تا بتوان استفاده از این داده ها را آسان کرد.همچنین بخش زیادی از این داده ها بدون استفاده می ماند که داده های سیاه عنوان میشود.که در ادامه بیشر در مورد انواع داده ها ، داده های سیاه و روش های روشن سازی صحبت خواهیم کرد.

۱-۱ انفجار اطلاعات

امروزه با پیشرفت و گسترش تکنولوژی ارتباطات نسبت به گذشته آسان تر شده است و افراد در همه جای دنیا با استفاده از ابزارهای ارتباطی فراوان قادر به برقراری ارتباطات صوتی وتصویری و متنی هستند.مخابرات، شبکه های اجتماعی، انواع برنامه های کاربردی ارتباطی و پیام رسانه ها، ابزاری در این دسته اند. شبکه های اجتماعی علاوه بر ایجاد ارتباطات دوستانه، محلی برای برقراری تجمع ها و گروه هایی متشکل از افرادی با تفکرات و عقاید یا مشاغل یکسان و همین طور محلی برای به اشتراک گذاری اخبار و اطلاعات روز دنیا است. تکنولوژی بستری فراهم کرده که هر فرد بتواند یک منبع تولید محتوا باشد. این اطلاعات تولید شده می تواند محتوایی آموزشی داشته باشد. مانند فیلم ها و دستور العمل های آموزشی در حوزه های مختلف علمی و هنری، محتوایی فرهنگی را به اشتراک بگذارد. مانند صفحات اجتماعی و وبلاگ های معرفی و نقد فیلم [۱]و یا معرفی و برگزاری دوره های کتابخوانی و انواع رویداد های فرهنگی[2]، در حوزه تبلیغات فعالیت کند یا حتی اطلاعاتی غیر مفید و آسیب زننده [۳]را منتشر کند. این اطلاعات شامل الگوها و اطلاعات ارزشمندی است که با جمع آوری، مدیریت و سازماندهی مناسب می تواند در حوزه های مختلف مانند انواع بیزینس ها [٤]در سطوح مختلف، سازمان های اطلاعاتی و امنیتی و حتی گاهی تصمیمات کلان اقتصادی کارا باشد.



شکل ۱- ۱ مثالی از انفجار داده[۵]

علاوه بر اینها بخش دیگری از اطلاعات دیجیتال اطلاعاتی است که در خود ادارات و سازمانها و شرکت های دولتی و خصوصی تولید می شود[۱] مثلا اطلاعات پرسنل شرکت ها و سازمانهای بزرگ که سالها در بایگانی شرکت ها باقی می ماند.گزارش گیری ها و اطلاعاتی که روند رشد شرکت ها را نشان می دهد اطلاعات کاربران در سازمان ها و مراکز مختلف مانند شرکت های بیمه، بیمارستان ها، آزمایشگاه ها، اداره ثبت و آمار گیری، دانشگاهها و موسسات آموزشی ، انواع ثبت نام های مجازی و در حوزه های پر اهمیت تر تراکنش های مالی اطلاعاتی هستند که در شرکت ها باقی می مانند.

توضیحات بیان شده در بالا نشان می دهد گسترش تکنولوژی همانقدر که مفید و کار آمد بوده است و در راستای حل مشکلات ما گام های بلندی برداشته است مشکلات جدیدی را نیز ایجاد کرده است.

حجم عظیمی از کلان داده ها در فرمت های مختلف برای مقاصد متفاوت تولید شده اند. این حجم عظیم چالش هایی را برای ما ایجاد می کنند. [۶]جمع آوری، نگهداری[۱]، پاکسازی[۲]، یکپارچه سازی[۳]، مدیریت و سازماندهی و هر پردازشی که برای رد شدن از این مراحل باید روی این حجم عظیم داده ها انجام داد. هنگام مواجه با داده های با اهمیت بالاتر مانند داده های امنیتی و مالی یا داده هایی که متعلق به افراد جامعه است و برای آنها از اهمیت خاصی برخوردار است [۷]گاها هویتشان را فاش میکند یا انتشار آن باعث بروز آسیب می شود باید بحث امنیت اطلاعات و حریم خصوصی افراد و حتی مشکلات حقوقی موجود در این زمینه را نیز در نظر گرفت.

۱-۲ چالش داده های بزرگ

هدف از ذخیره‌سازی داده‌ها نظم دادن به دانش خودمان است. پیداکردن الگوهای[۸] مشابه در میان این مجموعه‌ی داده‌ها ، می‌تواند قدرت پیش‌بینی را افزایش دهد. هرچه این داده‌ها بیشتر باشند و ارتباط آن‌ها دقیق‌تر باشد، یافتن این الگوها آسان‌تر خواهد شد. از طرفی دیگر در دنیای کنونی و افزایش جمعیت، پیشرفت دانش، گسترده‌شدن وسایل ارتباطی و علاقه به ارتباطات جدیدتر و پیچیده‌تر، باعث شده است که حجم[۴]، تنوع[۵] و سرعت[۶] داده‌هایی که در حال ذخیره‌ی آن‌ها هستیم به شدت افزایش یابد. [۹]

برای رد شدن از مراحل مختلف مانند:جمع آوری، نگهداری، پاکسازی، یکپارچه سازی، مدیریت و سازماندهی داده ها و اطلاعات دیجیتال روش های مختلف و متنوعی در حوزه های مختلف علم کامپیوتر اعم از داده کاوی[۷]، شبکه های عصبی[۸]، یادگیری ماشین[۹]، هوش مصنوعی[۱۰]، شناسایی الگو[۱۱] و الگوریتم های تکاملی[۱۲] ابداع شده است که ما

---

۱ Data curation

۲ Data cleaning

۳ Data integration

۴ volume

۵ variety

۶ velocity

۷ Data mining

۸ Neural networks

۹ Machine learning

۱۰ Artificial intelligence

۱۱ Pattern recognition

۱۲ Evolutionary Algorithms

از هر کدام به فراخور نوع داده ورودی، نوع داده خروجی، نوع انتشار، حجم داده ومحدودیت های موجود بهره می جوییم.[۱۰]

در حوزه های مطرح شده هنوز کمبود ها و نقایص زیادی وجود دارد خیلی از الگوریتم ها محدودیت هایی دارند که رعایت آنها خیلی برای مخاطب و کاربر خوشایند نیست و گاها دردسر ساز است . علاوه بر این مسائل و مشکلات مطرح شده در رویا رویی با حجم کلان داده چند برابر شده و مشکلات جدیدی نیز به آنها اضافه میشود. با گسترس روز افزون اطلاعات، چالش کلان داده به چالش ها و مشکلات قبلی اضافه می شود و برای اعمال الگوریتم های اثبات شده و معرفی شده روی حجم کلانی از اطلاعات نیازمند روش ها، الگوریتم ها، و ابزارهای خاصی هستیم.با معرفی این حوزه به عنوان یک چالش جدید در علم کامپیوتر و تحقیق و سرمایه گذاری روی آن روش ها و الگوریتم های تازه ای ابداع شد و در ادامه ابزارهایی تولید شد تا با استفاده از آنها بتوان حجم عظیم داده ها را با کارایی مناسبی پردازش کرد.[۱۱]

این بعد از علم کامپیوتر تقریبا حوزه ی جدیدی تلقی می شود و محققان و دانشمندان در سراسر جهان کماکان مشغول کار روی آن هستند.

۱-۳ تعریف مسئله

انفجار اطلاعات چالش داده های بزرگ را ایجاد میکند و داده های تیره به عنوان بخش عظیمی از داده های بزرگ چالش جدیدی را به ارمغان می آورد یکی از تعاریفی که برای داده های تاریک[۱۳] مطرح شده است این است که هر داده ی بدون فرا داده[۱۴] و بدون برچسبی را بتوان داده تاریک در نظر گرفت. داده های بدون ساختار بخش اعظم داده های تاریک را تشکیل می دهند. داده هایی که توسط محققان اندیشمندان و یا مردم عادی تولید می شوند ولی هیچ گاه به مصرف عموم مردم نمی رسند. [۱]در این مقاله به این نوع داده ها به عنوان یک چالش علمی در مسیر پیشرفت نگاه شده است و بر آن است تا راه حلی برای حل این مشکل بیابد.

راه حل های موجود از دو منظر سعی در حل این چالش دارند یکی با تلاش برای عدم ایجاد داده های تاریک و دیگری با تلاش برای روشن سازی داده های تاریکی که هم اکنون موجود است.

---

[۱۳] Dark data

[۱۴] Meta data

۱-۴ اهداف پژوهش

سوالی که در اینجا مطرح می شود این است که آیا حل این چالش مشکلی حیاتی است؟ جواب سوال مثبت است داده های تاریک مثال دریاچه های بی حاصلی است که اگه تجمیع و پاکسازی شود دریایی از اطلاعات ارزشمند را به ارمغان می آورد.[۱] که این اطلاعات هم به لحاظ علمی و هم به لحاظ اقتصادی بسیار ارزشمند و حائز اهمیت است. هدف اصلی از بررسی این داده ها و تلاش برای روشن سازی و بکار گیری آنها، رسیدن به دریای بزرگ اطلاعات یکپارچه، روشن و در دسترس است که به رشد علم در حوزه ها و ابعاد مختلف آن، کمک شایان توجهی می کند. بی توجهی به این اطلاعات، بی توجهی به اطلاعاتی است که اندیشمندان و محققان برای به دست آوردن آنها تلاش بسیاری کرده اند. در واقع هدف اصلی از پرداختن به این بعد از چالش کلان داده، تلاش برای گسترش مرزهای علم و دانش با استفاده از دانش نهان موجود است.در واقع این حوزه از علم در تلاش است تا دانش نهانی را هویدا سازد که اگر کماکان در غبار بماند حجم عظیمی از اطلاعات را به همراه خود مدفون کرده و نابود می سازد. [۶]

۱-۵ ضرورت ها و کاربردها

یکی از دلایل پرداختن به داده های تاریک به عنوان یک چالش اساسی عدم تولید دوباره بعضی مجموعه داده ها است این داده های حیاتی اگر غبار روبی و جمع آوری نشوند بطور کامل از بین رفته اند و امکان پیشرفت از بعدی از علم گرفته می شود.در این حوزه می توان به داده های مربوط به محیط زیست یا داده های ژنتیکی اشاره کرد[۶]

۱-۶ خلاصه

چنانچه ذکر شد کلان داده ها یا همان داده های بزرگ با افزایش روز افزون داده و محتوا ایجاد شد .پیشرفت تکنولوژی به ما امکان ایجاد و ذخیره حجم زیادی از داده ها را می داد. اما برای ما مشکلاتی هم ایجاد میکرد .زیرا وقتی هر کس مستقلا و بدون نظارتی بتواند محتوا تولید و انتشار دهد نگهداری، مدیریت، سازماندهی و پردازش این داده ها سخت می شود .همچنین این داده ها بدون هیچ فرمت خاص تولید شده در نتیجه قبل از استفاده و مدیریت باید بتوان آنها را یکپارچه کرد.

در این سمینار، کارهای انجام شده در تجزیه و تحلیل انواع داده در سه دسته مورد بررسی قرار گرفته است که عبارتند از:

۱- چالش کلان داده ها
۲- چالش داده های تاریک
۳- روشن کردن داده های تاریک

فصل دوم به چالش داده های تاریک پرداخته است . فصل سوم انواع روش ها و کارهای انجام شده در این حوزه را شرح داده شده است و فصل چهارم نتیجه گیری نهایی از کل اطلاعات مطرح شده است.

فصل دوم : داده های تاریک

چالش جدید مطرح شده در بحث کلان داده ها مبحث داده های تاریک است . سوال مطرح شده این است که آیا تمام داده های تولید شده داده های تاریکند؟ داده های تاریک چند درصد از کل کلان داده ها را در بر میگیرد؟آیا ما قادر به استفاده از کل داده ها ی تولید شده توسط افراد و منابع مختلف هستیم؟آیا کل داده های تولید شده توسط یک شرکت در اتخاذ تصمیمات کلان اقتصادی آن شرکت نقش دارد؟آیا ما توان نگهداری تمام داده های تولید شده را داریم؟آیا نگهداری تمام داده های تولید شده کاری درست و منطقی است؟در این فصل تعریفی از داده های تاریک ارئه شده است و راجع به افراد، سیستم ها و شرکت هایی که از پرداختن به مبحث داده های تاریک به عنوان یک چالش سود می برند صحبت شده است و در ادامه به یافتن پاسخ این سوالات پرداخته شده است.

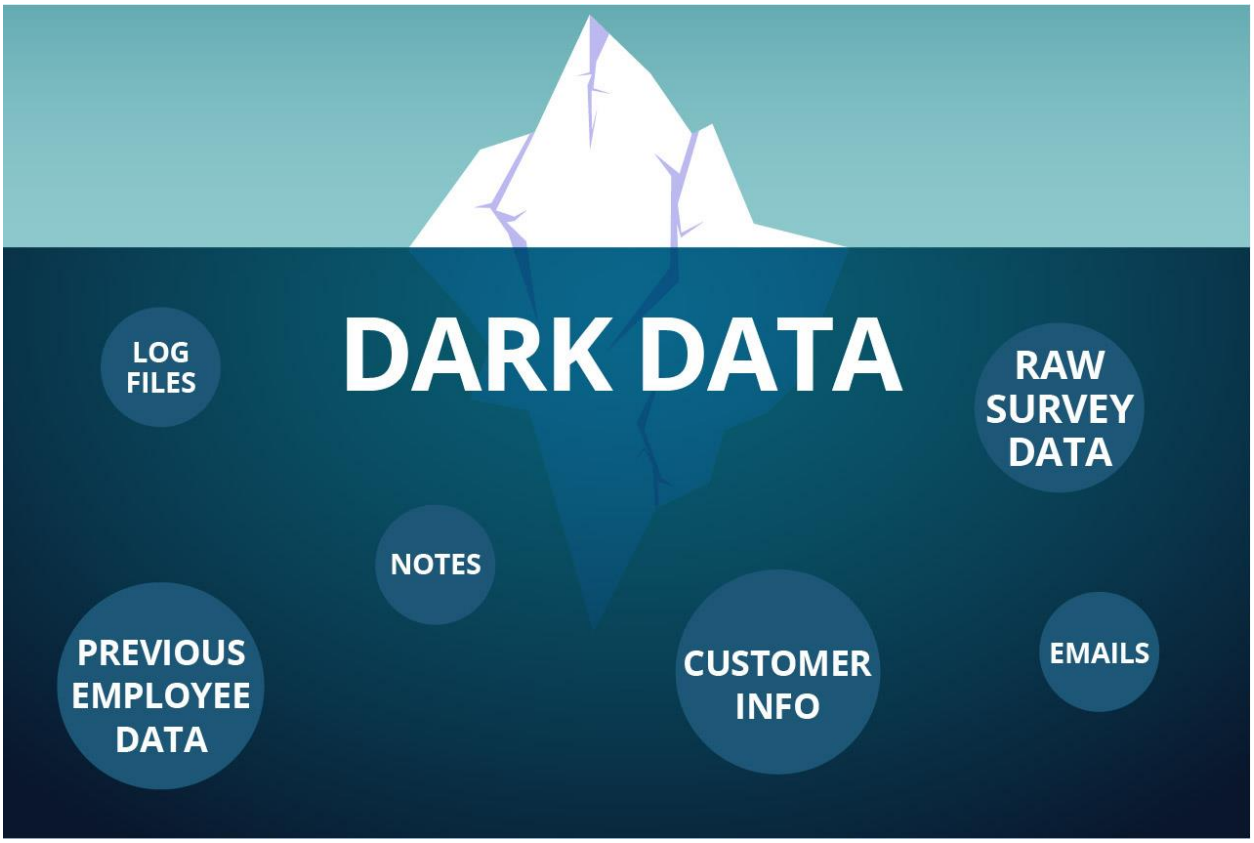
۲-۱ مقدمه ای بر داده های تاریک

داده تاریک، یکی از تازه‌ترین مباحث در حوزه داده‌های بزرگ و تجزیه و تحلیل آن‌ها است و تلاش می‌کند به پرسش بالا پاسخ دهد؛ مبحثی که اگر به درستی شناخته و به کار گرفته شود، نه تنها می‌تواند منبع درآمد مناسبی برای سازمان‌ها و شرکت‌های بزرگ باشد؛ بلکه از یک سو می‌تواند باعث تحرک و رونق شرکت‌های فنی مرتبط با بیگ دیتا، پردازش و تجزیه‌وتحلیل، داده‌کاوی و ذخیره‌سازی داده‌ها شود و از سوی دیگر، می‌تواند اطلاعات بسیار مناسبی را برای شناخت رفتار کاربران در اندازه‌های بزرگ، در اختیار اقتصاددانان، جامعه شناسان، روانشناسان اجتماعی و برنامه‌ریزان شهری قرار دهد.

۲-۲ تعریف داده های تاریک

تعریف موسسه گارتنر[۱] از دارک دیتا چنین است: «اطلاعاتی که یک سازمان در طول فعالیت عادی خود، گردآوری، پردازش و ذخیره‌سازی کرده است و جزیی از دارایی‌های آن به حساب می‌آید؛ اما نتوانسته است برای مقاصد دیگری از آن‌ها استفاده کند.عده‌ای در تعریف دارک دیتا، بر نقش آن در تصمیم‌گیری‌های سازمان‌ها و راهبردهای آنان در آینده، تاکید می کنند.سازمان‌های بزرگی مانند تامین اجتماعی و سایر موسسات بیمه‌ای، سازمان فنی حرفه‌ای، آموزش و پرورش، بانک‌ها، شرکت های ارائه‌کننده خدمات تلفن ثابت و همراه و تعداد زیادی از موسسات دیگر با این مبحث مرتبط هستند. یکی از این سازمان‌ها را در نظر بگیرید. این سازمان در جهت انجام کارهای عادی خود در طول یک سال، با ده‌ها و بلکه صدها هزار انسان سر و کار دارد و به نوعی، اطلاعات آن‌ها را در جایی ذخیره می‌کند؛ اما در بسیاری از موارد، به غیر از همان استفاده اولیه از این اطلاعات، هیچگونه استفاده دیگری از این داده‌ها صورت نمی گیرد اگرچه بخشی از این داده شاید، جزء حریم خصوصی مردم باشد و استفاده از آن‌ها چه به لحاظ قانونی و چه به لحاظ اخلاقی، مجاز نباشد؛ اما بخش‌های دیگری از آن‌ها می‌تواند، در مقاصد پژوهشی و بررسی‌های اجتماعی و راهبردهای کلان اقتصادی، مورد استفاده قرار گیرد. یکی از مشکلات مربوط به این داده‌ها، ذخیره‌سازی و امن نگه داشتن آن‌هاست که هزینه بالایی طلب می‌کند و این در حالی است که در بسیاری از موارد، هنوز ارزش این داده ها مشخص نشده است .دارک دیتا، نوعا بدون ساختار، بدون برچسب و دست‌نخورده، در درون انباره‌های ذخیره‌سازی یافت می‌شود و عموما تجزیه و تحلیل نشده است. این داده‌ها شبیه کلان داده ها هستند؛ با این تفاوت که ارزش آن ها عمدتا توسط سازمان یا مدیران آی‌تی، مورد غفلت قرار گرفته است. اغلب داده‌های تاریک، به‌گونه‌ای ذخیره شده‌اند که برای تجزیه و تحلیل دشوار، پیچیده و پرهزینه هستند؛ همچنین این داده‌ها می‌توانند اطلاعاتی را شامل شوند که توسط خود شرکت تهیه نشده‌اند و خارج از سازمان، توسط مشتریان یا شرکا ذخیره

شده‌اند.با رشد نمایی داده‌های ساخت‌یافته، نیمه‌ساخت‌یافته و بدون ساختار در سازمان‌ها، دارک دیتا به معنای داده‌های عملیاتی در نظر گرفته می‌شود که می‌تواند قابلیت تجزیه وتحلیل را پیدا کند؛ اگر شرکت‌ها ارزش این داده‌ها را بدانند، می‌توانند از آن‌ها به عنوان فرصتی برای افزایش درآمد یا کاهش هزینه‌های داخلی خود، استفاده کنند. بعضی از داده‌هایی که می‌توانند در این دسته قرار بگیرند، شامل این موارد هستند: فایل‌های لاگ‌سرور که کلیدهای رفتاری بازدیدکنندگان وب‌سایت‌ها را ارائه می‌دهند، جزییات ضبط شده تماس‌های تلفنی که احساسات و عواطف مشتریان را نشان می‌دهد یا داده‌های مربوط به موقعیت‌های مکانی دارندگان موبایل، که الگوهای ترافیکی را آشکار می‌کنند؛



شکل۲- ۱مثالی از داده های تاریک[۱۲]

همچنین دارک دیتا می‌تواند برای توصیف داده‌هایی به کار رود که مدت‌هاست در دسترس نیستند؛ زیرا روی وسایلی ذخیره شده‌اند که منسوخ شده‌اند[۱].

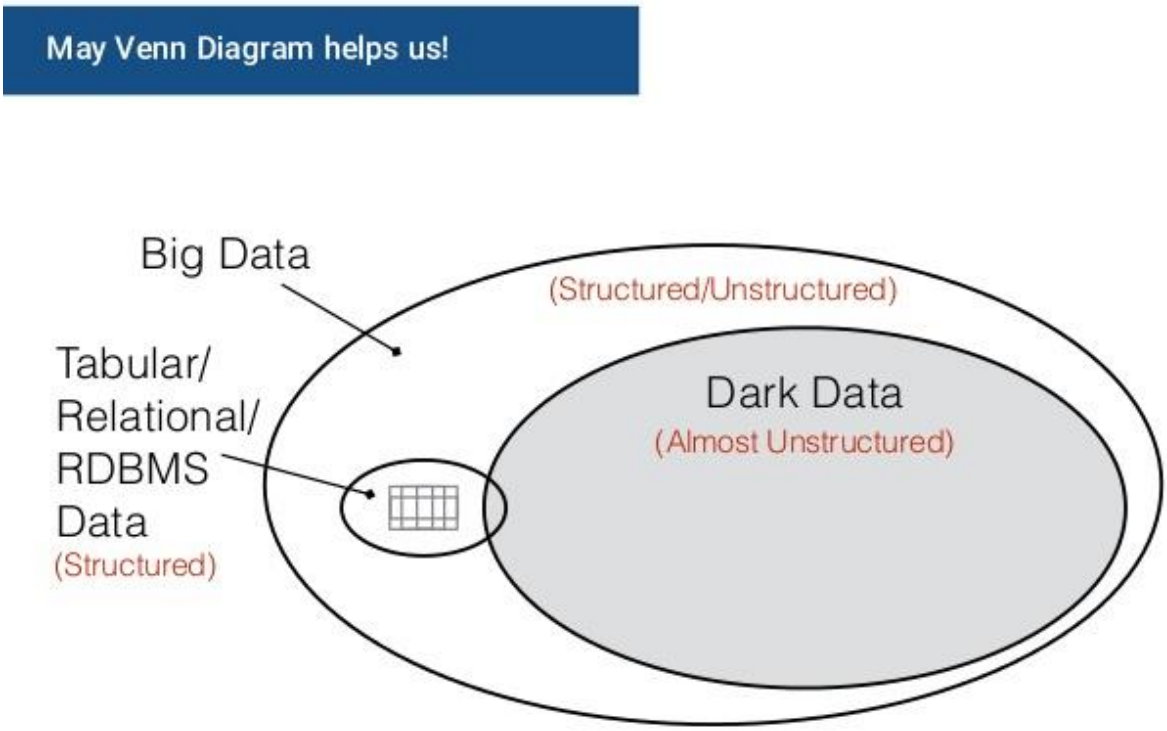
۲-۳ انواع داده های تاریک

داده های تاریک انواع مختلفی دارند که در ادامه راجع به آنها صحبت شده است.

انواع داده های تاریک:

۱. داه‌هایی که به تازگی جمع‌آوری نشده‌اند.
۲. داده‌هایی که جمع‌آوری شده‌اند؛ اما دسترسی به آن‌ها در زمان و در جای مناسب دشوار است.
۳. داده‌هایی که جمع‌آوری شده‌اند و در دسترس هستند، اما هنوز پردازش نشده‌اند.

شاید بتوان به این سه دسته از داده‌ها، نوع چهارمی را نیز افزود که شامل داده‌هایی می‌شود که سازمان‌ها هر روز آن‌ها را تولید می‌کنند؛ اما در جایی ذخیره نمی‌کنند.داده تاریک برخلاف ماده تاریک، این ظرفیت را دارد که پرتو نوری برآن افکنده شود و سرمایه‌گذاری مجددی روی آن انجام شود؛ در واقع موضوع اصلی این است که چگونه می‌توان با استفاده از روش‌های علمی و بر اساس روش فایده- هزینه، پیچیدگی‌ها و رمز و راز اطراف داده تاریک را حذف کرد و آن را برای استفاده و سرمایه‌گذاری مجدد آماده کرد[۱۳].



شکل۲- ۲ تفاوت بین داده های تاریک و کلان داده ها[۱۴]

۲-۴ چه کسانی با داده های تاریک سر و کار دارند

۱. شرکت‌ها و سازمان‌هایی که با حجم زیادی از اطلاعات کارمندان، مشتریان و سرمایه‌گذاران در ارتباط هستند؛ مانند شرکت‌های بیمه، خودروسازی، لیزینگ، هولدینگ، بانک‌ها، آموزش و پرورش، دانشگاه‌های بزرگ و وزارتخانه‌ها، این شرکت ها موظفند اطلاعات پرسنل و ارباب رجوع ها را ذخیره و نگهداری کنند و چون شرکت های بزرگی هستند در نتیجه تعداد افرادی که باید اطلاعات متفاوتی از آنها را ذخیره کنند زیاد است در نتیجه حجم داده ی ذخیره شده انقدر بالاست که در دسته کلان داده ها قرار می گیرد.اما گاها به دلیل همین حجم بالا از بخش بزرگی از این داده های ذخیره شده در پردازش ها استفاده نمی شود و یا اینکه یک سری از موجودیت ها یا افراد از سیستم خارج شده اند ولی اطلاعات آنها کماکان در سیستم وجود دارد این اطلاعات همان داده های تاریک هستند.

۲. . وب سایت‌هایی که در کار خرید فروش کالا و خدمات هستند یا وب‌سایت‌های خبری و محتوایی که با مخاطبان زیادی سروکار دارند.اطلاعات موجود در این وب سایت ها با نرخ بالایی روز به روز در حال افزایش است اما اگه ساختار محتواهای منتشر شده در این وب سایت ها یکپارچه نباشند آیا میتوان از آنها بهره لازم را برد و و از آنها به عنوان داده ورودی سیستم های داده کاوی استفاده کرد؟پاسخ منفی است.این داده ها نیز در دسته ی داده های تاریک جای می گیرند.

۳. شرکت‌های کامپیوتری که در کار هوش مصنوعی، داده کاوی،کلان داده و ذخیره‌سازی داده‌ها هستند.

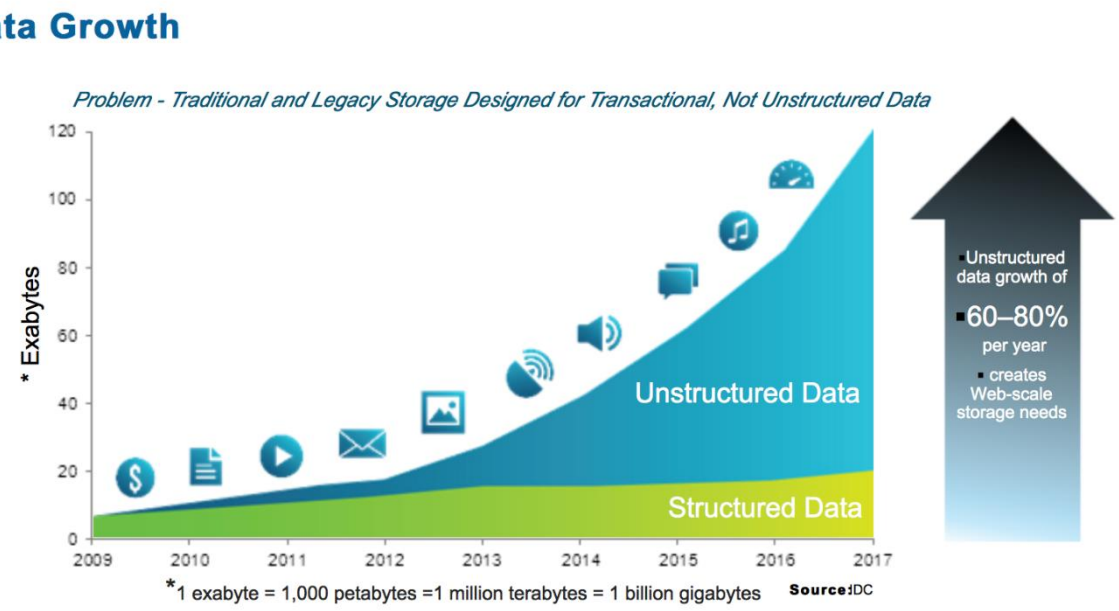
۴. شرکت‌هایی که در کار ساخت و یا ارائه دیتاسنتر و دیگر ابزارهای شبکه هستند .

۵. شرکت‌هایی که ارائه‌دهنده خدمات تلفن همراه، اینترنت موبایل، هاستینگ، دامنه و ... هستند

۶. متخصصان هوش مصنوعی، بیگ‌دیتا، شبکه، مدیران آی‌تی شرکت‌ها و سازمان‌های بزرگ.

۲-۵ ظرفیت داده های تاریک

کارشناسان داده‌کاوی اعتقاد دارند که برای یافتن یک تصویر جامع و کامل از یک مشتری، باید به سراغ معدن داده‌های تاریک او رفت. اما، این کار، چندان که به نظر می‌رسد ساده نیست. تقریبا هیچکس در یک شرکت نمی‌داند که با این داده‌ها چه باید بکند و یا حتی آن را چگونه تحلیل کند. چرا که این داده‌ها معمولا به روشی درست و کاربردی جمع‌آوری و ذخیره‌سازی نشده‌اند. اغلب آن‌ها به‌صورت خام هستند. این داده‌ها فهرست‌شده و در حال استفاده نیستند. حتی بسیاری از سازمان‌ها از وجود آن‌ها آگاه نیستند. اما در مجموع این کارشناسان معتقدند که هر گونه اطلاعاتی که به شما اجازه دهد که بین خود و مشتری‌تان و یا میان مشتریان‌تان ارتباط برقرار کنید، حتما از ظرفیت بالایی برخوردار خواهد بود. داده تاریک به کسب وکارها اجازه می‌دهد که تصویری دقیق از مشتریان خود کسب کنند تا بتوانند بهترین پیشنهادها را به آن‌ها ارایه دهند. این امر موجب رونق کسب‌وکارها و ارتباط بهتر میان مشتری و شرکت خواهد شد.



شکل۲- ۳ شیوه توزیع داده های ساخت یافته و بدون ساختار[۱۵]

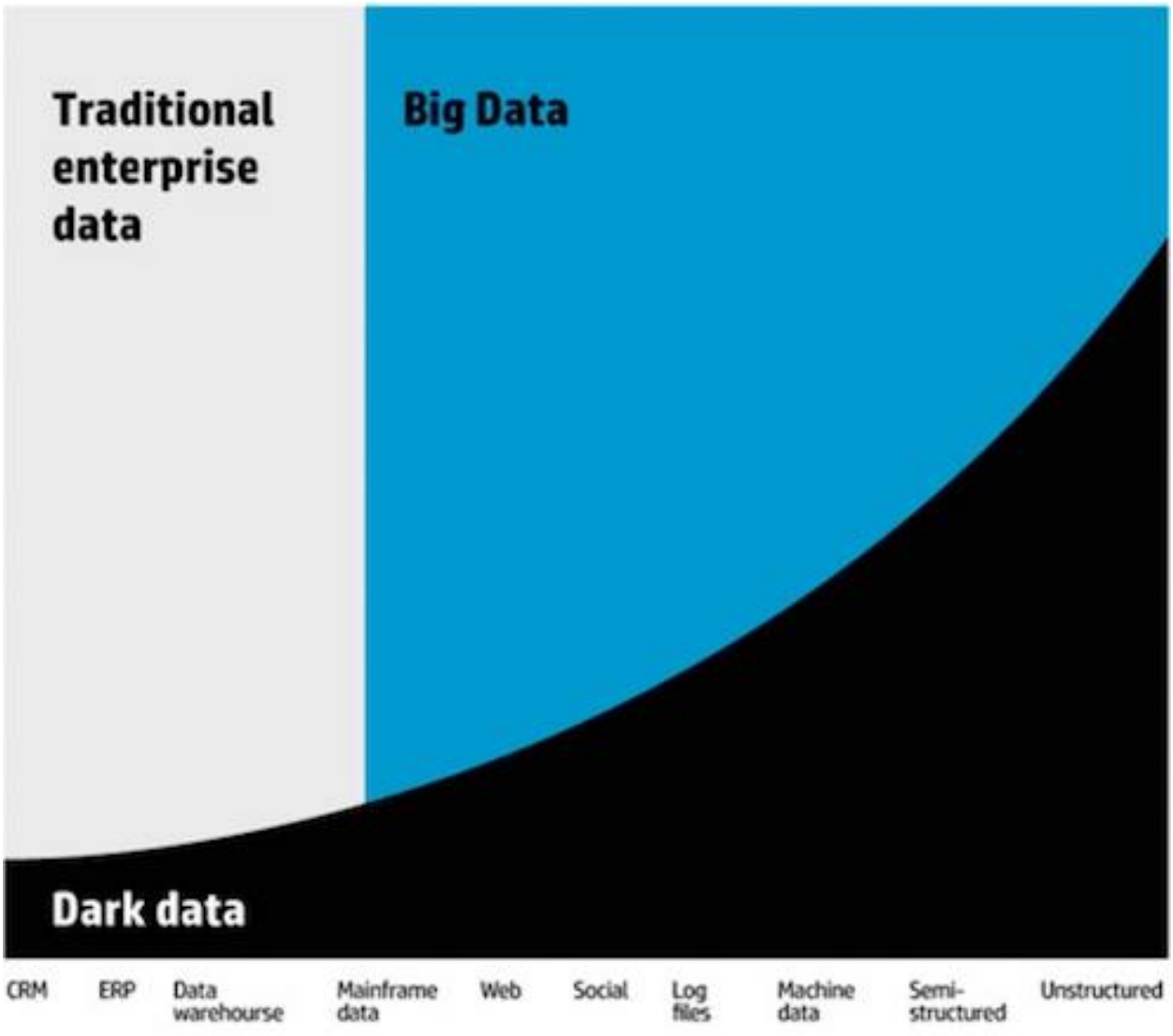
۲-۶ آینده ی داده های تاریک

شاید برای شرکت‌هایی که تازه شروع به فعالیت می‌کنند، در کوتاه مدت، موضوع داده‌های تاریک خیلی مهم نباشد. آن‌ها داده‌های تازه خود را تولید می‌کنند؛ اما در گذر زمان یک روز متوجه می‌شوند که انبوهی از داده‌ها که نمی‌دانند چیست، در انباره‌های خود ذخیره کرده‌اند و نمی‌دانند با آن‌ها چه کنند. اینجاست که باید به سراغ افراد و شرکت‌هایی بروند که این توانایی را دارند که بتوانند از داده‌های تاریک ارزش بیافرینند. بسیاری از متخصصان حوزه کلان داده و داده‌کاوی، باید خود را برای این حوزه جدید آماده کنند. حوزه‌ای که این بار چیزهای باارزش را، از دل تاریکی‌ها بیرون می‌کشد؛ البته تعدادی از صاحب نظران این حوزه، پرداختن به داده‌های تاریک را خطرناک می‌دانند. عمده این خطرات، مشکلات قانونی و آلودگی اطلاعاتی و مسائل اخلاقی هستند؛ بدیهی است، بحث‌هایی که در این گزارش در جهت استفاده از داده تاریک به آن‌ها اشاره شده است، با در نظر داشتن مباحث قانونی و اخلاقی است. داده‌هایی از این دست، جزء دارایی‌های هر سازمان به حساب می‌آیند و تصمیم گیری در مورد استفاده یا عدم استفاده از آن‌ها، تنها برعهده مالک آن و با رعایت موازین قانونی و اخلاقی، مجاز است[۱۳].

۲-۷ تولید داده های تاریک

یکی از خروجی های اصلی شرکت علمی، داده ها است، اما بسیاری از موسسات مانند کتابخانه هایی که مسئول حفظ و انتشار خروجی های علمی هستند، این نوع خروجی را نادیده می گیرند. این گزارش بر روی یک کلاس ویژه از داده ها تمرکز دارد که داده های تاریک نامیده می شود. "داده های تاریک" تقریبا نامرئی برای دانشمندان و سایر کاربران بالقوه است و بیشتر احتمال دارد که استفاده نشده و در نهایت از دست برود. این گزارش در مورد چگونگی استفاده از مفاهیم اقتصاد طولانی مدت برای درک راه حل های بالقوه بحث می کند. در ادامه توضیح داده می شود که چرا این داده ها برای پیشرفت علمی حیاتی اند، و همچنین برخی موانع اجتماعی و فنی را شرح داده می شود. اکثریت کارهایی که توسط دانشمندان انجام می شود، در پروژه های نسبتا کوچک انجام می شود که یک پژوهشگر برجسته به همراه دو یا سه دانشجوی فارغ التحصیل به صورت نیمه وقت روی یک یا چند موضوع و پروژه ی کوچک کار میکنند. محصول خام این تلاشها، داده های علمی و اطلاعاتی است که پایه و اساس تمام نظریه های علمی را تشکیل می دهد. در حالی که هزینه های زیادی به جمع آوری، نگهداری و استفاده مجدد از داده ها در پروژه های بسیار بزرگ اختصاص داده شده است، نسبتا توجه کمتری به داده هایی که توسط اکثر دانشمندان تولید می شود، اختصاص داده شده است. ساختارهای اجتماعی جدید و پیشرفت های فنی می تواند به میزان قابل توجهی در دسترس بودن و ارزش داده های دانشمندان مستقل در تحقیقات مربوطه را افزایش دهد.

پروژه های علمی را در امتداد یک محور از بزرگ تا کوچک سازماندهی می شوند. پروژه های بسیار بزرگ که از ده ها یا بیشتر دانشمند از آن حمایت می کنند در سمت چپ محور قرار می گیرند و حجم زیادی از داده ها را تولید می کنند و پروژه های کوچکتر در سمت راست قرار می گیرند. اگر کل داده ها ی علمی خروجی از کل پروژه های علمی و تحقیقاتی تولید شده تا به امروز را توسط گراف ها مدل سازی کنیم سمت چپ گراف، تعداد نود های کمتر با ارزش خیلی بیشتر و سمت راست گراف، میلیاردها یا حتی بیشتر نود، با ارزش هایی بسیار کم قرار می گیرد. در این مدل سازی نود ها، پروژه های تعریف شده هستند که با توجه بزرگی پروژه، اعتبار و شهرت محققین ای که مسئولیت آن را برعهده دارند، به هزینه هایی که از جانب موسسات آموزشی، دولتی، و گاها خصوصی برای پیش بردشان در نظر گرفته می شود و تعداد و ارزش افرادی که خواهان تعریف پروژه هستند ارزش دهی شده و وزن دار میشوند.و یال های گراف ارتباط بین پروژه های مختلف علمی است که در هر حوزه ای معیار های تعریف مشخصی دارد.داده های تولید شده توسط پروژه های بزرگ و مهم که در سمت چپ گراف قرار گرفته ساختار یافته تر است زیرا هزینه ی زیادی برای پیش برد این پروژه ها و نگهداری از دیتا های تولید شده انجام می شود. داده های تولید شده یکپارچه هستند و توسط مخازن بزرگ ذخیره سازی اطلاعات ذخیره می شوند این داده ها توسط صفحات وب، کتابخانه ها و دیگر مراکز نشر اطلاعات علمی در اختیار کاربران قرار می گیرند [۱]



شکل ۳- ۱نحوه ی توزیع شدن داده های تاریک[۱]

عملا داده های تولید شده توسط محققین و دانشمندان علوم مختلف به دو دسته ی داده های مفید و داده های غیر مفید تقسیم می شوند داده های غیر مفید، نتیجه آزمایش هایی هستند به درستی اجرا نشده اند. داده هایی که عملا برای ما منفعتی ندارند. ولی آزمایشاتی که به درستی و در شرایطی کنترل شده انجام شوند چه به نتیجه ی دلخواه برسند، چه نرسند، داده هایی مفید تولید می کنند .حتی اگر آزمایشی به نتیجه نرسد، باز هم نتیاج آن برای دانشمندان و دانش پژوهان آن حوزه ی خاص اطلاعاتی مفید و کاربردی است و به آنها کمک می کند که در راه حصول به نتیجه نهایی مسیر های بهتری را انتخاب کنند. اما ما در عمل تنها داده های مفیدی که حاصل آزمایشات موفق هستند

را منتشر می کنیم و حتی گاهی به دلیل حجم زیاد داده های تولیدی، از بین داده های مفید و نتایج مثبت نیز تعدادی را به نمایندگی انتخاب می کنیم.داده های غیر مفید، داده هایی که نتایج آزمایشات موفق نیست و حتی گاهی بعضی از داده های مفیدی که نتیجه ی آزمایشات موفق است در دسترس عموم قرار نمی گیرد و در دسته ی داده های تاریک قرار می گیرد.به داده های تاریک، داده های تیره، داده های کثیف، داده های مه آلود یا داده های غبار گرفته نیز گفته می شود.اما توضیحات بالا صرفا برای سمت راست گراف مدل ما بود که بخشی از داده های آن روشن و در دسترس و بخشی از آن تیره و در غبار بود. برای سمت راست گراف که شامل تعداد بی شماری نود با ارزش کمتر است چه اتفاقی رخ میدهد؟ پروژه های تعریف شده در سمت راست گراف مدل معمولا پروژه هایی است که توسط محققین، دانشمندان، اساتید دانشگاه، و فارغ التحصیلان مستعد رشته های مختلف بصورت مستقل تعریف می شود. حتی نتایج مثبت ایجاد شده توسط آزمایشات موفق نیز در این نقطه از گراف شانس زیادی برای روشن و در دسترس عموم قرار گرفتن ندارند، چرا که از جانب محققان مشهور و موسسات و مجموعه های خصوصی و دولتی تاثیر گذار حمایت نمی شوند، در فرمت های خاص تعریف شده قرا ر نمی گیرند، با لینک های تاثیر گذار ارتباط برقرار نمی کنند و قابلیت و شرایط لازم برای تامین هزینه ی کافی برای نگهداری و نشر اصولی داده ها را ندارند. حتی گاهی انگیزه ی کافی برای این کار هم ندارند در نتیجه اکثریت داده های تولید شده در این نقطه تاریک و در غبار باقی می ماند به نشر عمومی نمی رسند و در دسترس مشتاقان قرار نمی گیرد و در مکان های مبهم (مثلا کشوی میز یا بلاگ های شخصی محققین) باقی می ماند تا طی گذر زمان از ارزش علمی آن کاسته شود، فراموش شود، دیگر قابل خواندن نباشد (چون دستگاه ها یی که قادر به خواندن آن هستند کنار گذاشته شده- اندیا محققین و متخصصینی که قادر به توصیف آن هستند دیگر در قید حیاط نیستند)و در نهایت از بین برود.

در این گزارش، از اصطلاح داده های تاریک برای اشاره به هر گونه اطلاعاتی که به آسانی توسط کاربران بالقوه پیدا نمی شود، استفاده شده است.[۱۳] داده های تیره ممکن است یافته های مثبت یا منفی تحقیق یا از علم "بزرگ" یا "کوچک" باشد. بسیاری از عوامل مورد نیاز به توجه بیشتر به پروژه ها و موضوعات مطرح شده در سمت راست گراف مدل می شود، از جمله تعداد روزافزون دانشمندان در سطح جهان و افزایش مقدار داده هایی که هر دانشمند می تواند با ابزارهای مدرن تولید کند. این رشد گسترده در جمع آوری داده ها به هیچ وجه تضمین نمی کند که داده هایی که در حال حاضر در دسترس هستند درآینده نیز قابل دسترسی باشند. یا داده های تولید شده در آینده بتواند به خوبی نگهداری و نشر شود. به همین منظور، باید ساختارهای اجتماعی جدید و ابزارهای مدیریت اطلاعات سریعا گسترش یابند و امکان مدیریت اطلاعات جدید را فراهم کند. تا حداقل نگرانی ها بابت از بین رفتن و بی فایده ماندن داده های تولید شده در آینده کمتر شود. ممکن است تنها تعداد کمی از دانشمندان در سراسر جهان وجود داشته باشند که بخواهند داده های تولید شده در رابطه با موضوعی خاص را مورد مطالعه و بررسی قرار دهند، اما هزاران نفر

از این مجموعه داده ها وجود دارد. دسترسی به این مجموعه داده ها می تواند تاثیر بسیار مهمی در علم داشته باشد. به نظر می رسد که علم تحول پذیر در سمت راست گراف مدل بیشتر از سمت چپ ساخت یافته- ی آن باشد. در بیشتر موارد، در پروژه های بزرگ در مقیاس بزرگ که داده های حجم بالا را تولید می کنند، سوالاتی که باید پاسخ داده شوند نسبتا خوب درک شده اند. پروژه های تعریف شده و داده های تولید شده در سمت راست گراف بستر مناسبی است برای پرورش ایده های علمی که قبلا هرگز روی آنها کار نشده است. پروژه های نامطمئن و خطرناک که اگر این امکان برایشان ایجاد شود که کمک مالی دریافت کنند می توانند بسیار موفق عمل کنند. داده های پنهان در سمت راست گراف و علم داده های تاریک محدودیت های اقتصادی و اجتماعی مشابهی دارد. داده های علمی ثروت هستند و دیدن این ثروت تقریبا غیر ممکن است. از آنجا که یافتن اطلاعات تاریک دشوار است، استفاده از آن کم است و به طور معمول از بین می رود. ما می توانیم بر اساس حجم داده هایی که تولید می کنند، پروژه های علمی را مرتب کنیم و نمودار کنیم و با توزیعی مناسب هزینه های لازم برای گسترش علم را در اختیار محققین و اندیشمندان قرار دهیم. ولی معمولا این توزیع به درستی اتفاق نمی افتد و چون بقیه قوانین طبیعی ناعادلانه قسمت می شود . این جا قانون هشتاد_بیست مطرح می شود و ۸۰ درصد از کل هزینه ی در نظر گرفته شده تنها به ۲۰ درصد از پروژه های تعریف شده تعلق می گیرد [۱].

پایگاه داده ها را می توان به دو دسته پایگاه داده های عمیق و پایگاه داده های گسترده تقسیم کرد پایگاه های اطلاعاتی "عمیق[۱۵]" در چند نوع داده متخصص هستند که جمع آوری داده های نسبتا همگن را موجب میشوند و اجازه می دهد تا ابزارهای جستجوی پیشرفته توسعه پیدا کنند. پایگاه داده های "گسترده" انواع بسیاری از داده ها را جمع آوری می کنند و توسعه ابزار برای مقابله با اطلاعات بسیار مشکل تر است.

ما گمان می کنیم که سمت چپ گراف مدل که ما آن را به عنوان سرگراف می شناسیم ممکن است تمایل داشته باشد که مجموعه داده های "عمیق" بیشتری داشته باشد، زیرا بیشتر این اطلاعات با ابزار اختصاصی جمع آوری می- شود. و سمت راست گراف که ما آن را به عنوان دم [۱۶] گراف می شناسیم به طور کلی بسیار ناهمگن است، اما دم ممکن است شامل بسیاری از پروژه هایی باشد که با انواع داده های مشابه کار می کنند. این اطلاعات پراکنده اما مشابه در دم نشان دهنده یک فرصت است.بعضی از حوزه های علم در حال حاضر برای جمع آوری داده ها و تبدیل آنها به مجموعه داده های "عمیق" سرمایه گذاری کرده اند. علم طولانی[۱۷] مدت به معنی سوالات علمی کوچک و یا حتی علم کوچک نیست. نتایج حاصل از پروژه های چندگانه در دم گراف مدل می تواند متخصصان را به اطلاعات واقعا

[۱۵] deep

[۱۶] Long tail of the science

[۱۷] Long science

بزرگ، دستاوردهای بزرگ و دانش انباشته شده برساند، در صورتی که اگر از آنها بدرستی استفاده شود. یک مثال نمونه از این نوع تحقیقات، پروژه های بیولوژی مولکولی است که به بانک ژن[۱۸] و بانک اطلاعات پروتئینی[۱۹] کمک می کند.

گردآوری این داده ها ممکن است منجر به پیشرفت هایی شود که قابل پیش بینی نیست. اطلاعات تیره در سراسر علم وجود دارد. در حالی که برخی ممکن است در واقع داده های نادرستی باشند که باید به دلیل اشتباه ها کنار گذاشته شود، برای جمع آوری داده های بالقوه مفید و غیرقابل استفاده، مقدار قابل توجهی از زمان و پول خرج می- شود. بسیاری از این اطلاعات تاریک در دم طولانی علم(سمت راست گراف مدل)، در آزمایشگاه های مستقل قرار دارند[۱].

۲-۸ درک داده های تاریک

به منظور مدیریت داده های تاریک دم علم بهتر است، این اطلاعات لازم درک شود. اکثر پیشنهادات در ابتدا در سر گراف قرار می گیرند، و سپس برخی به سمت دم می روند. اما این تلاشها به طور مستقیم به برخی از سوالات مربوط به داده های تاریک در دم نمی پردازد. برخی از سوالات مهم به شرح زیر است:

انتهای این گراف کجاست یا عمق این مخزن تاریک چقدر است؟

چه داده هایی در سر گراف و چه داده هایی در دم گراف تاریک هستند؟

معیار ارزش دهی به این داده های تاریک چیست؟

چه تفاوتی بین علمی که در سر گراف است با علمی که در دم ان است وجود دارد؟

به چه حجمی از دیتای تولیدی چه اندازه کمک مالی تغلق میگیرد؟

کدام یک از دیتا ها در پیشرفت و تغییر علم موثرند؟

تعدادی از روش ها می توانند برای درک داده های تاریک مورد استفاده قرار گیرند اما همه در نهایت مطالعه رفتارهای دانشمندان انفرادی است. اقتصاددانان و جامعه شناسان علوم درحال انجام نظرسنجی، مصاحبه و مطالعات نظارت هستند تا بدانند چگونه داده ها در رشته های علمی مختلف مورد پردازش و ذخیره قرار می گیرند. این اطلاعات به ما کمک می کند تا مکانیزم های بهتر برای حمایت از استفاده مجدد از داده های علمی گسترده تر را طراحی کنیم.

---

[۱۸] Genbank

[۱۹] PBD

۲-۹ خلاصه

در این فصل مفهوم ابتدایی داده های تاریک مطرح شده است و راجع به منابعی که داده های تاریک را ایجاد می کنند بحث شده است و در این حوزه مثال هایی از دنیای واقعی ارائه شده است.از ارزش و ظرفیت داده های تاریک سخن گفته شده است و اینکه اگر برای مدیریت و سازمان دهی آنها زمان و انرژی و هزینه صرف شود چقدر می توانند در حوزه های مختلف و برای گسترش علوم مختلف مفید و کاربردی باشند و در نهایت توضیح کوچکی راجع به آینده ی داده های تاریک داده شده است و چشم اندازی که خیلی قابل پیش بینی نیست.

فصل سوم: کارهای انجام شده در حوزه داده های تاریک

در بالا تعاریفی کلی از داده های تاریک ارائه شده است چرا داده های تاریک تولید می شوند؟آیا می توان تولید داده های تاریک را متوقف کرد ؟آیا میتوان داده های تاریک را به روشنایی آورد؟چه حجم از داده های تاریک می توانند به روشنایی آورده شوند؟روش های روشن کردن داده های تاریک به چه شکل است؟در ادامه به توضیح پاسخ سوالات بالا پرداخته می شود.

۳-۱ راه حل ها ی بلقوه سازمانی آوردن داده های تاریک به نور

موسسات موجود و در حال توسعه نقش حیاتی در بهبود دسترسی به داده های تاریک بازی می کنند. بعضی از راه-حل های مشابه برای مدیریت داده ها که برای سر گراف طراحی می شوند می توانند برای دسترسی به داده های موجود در دم استفاده شوند. در حقیقت، در برخی از رشته ها این فرآیند قبلا شروع شده است. یک راه حل این است که مراکز داده های علمی درمورد رشته های فردی ایجاد کنند. بسیاری از ابتکارات در سازمان های فدرال مانند NSF[۲۰]، ناساو NOAA[۲۱] اتفاق افتاده است ومحققان اصلی در حال حاضر به این مسائل پاسخ می دهند. برای مثال، NSF در سطح سازمانی، از ایجاد مرکز ملی تجزیه و تحلیل و سنتز زیست محیطی برای تأمین مسائل مربوط به داده ها حمایت کرد.

مأموریت NCEAS[۲۲] سه گانه است. اول، جستجوی الگوهای عمومی و اصول در داده های موجود و بررسی وضعیت دانش زیست محیطی دوم، سازماندهی و تولید اطلاعات زیست محیطی به شیوهای مفید برای محققان، مدیران منابع و سیاستگذاران برای حل مسائل مهم محیط زیست سوم، بر روش تحقیق زیست محیطی تأثیر می گذارد و فرهنگ سنتز، همکاری و به اشتراک گذاری داده ها را تبلیغ می کند.

کتابخانه ها در حال حاضر با مشکلاتی طولانی روبرو هستند. کتابخانه ها به طور فزاینده ای درنگهداری اطلاعات علمی نیز نقش دارند [۱٦] اما با چالش های فرهنگی و مالی مواجه هستند [۱۷]. در حالی که بسیاری از ابتکارات با تمرکز بر متن دیجیتالی آغاز شده است، این تجربه ها راه را برای مدیریت خروجی های علمی دیگر هموار کرده است. انجمن کتابخانه های تحقیقاتی در یک سری از کارگاه ها و نشریات در مورد مراقبت از داده ها شرکت کرده است.

بسیاری از کتابخانه ها در حال حاضر مخازن داده های سازمانی را ایجاد کرده اند، که گاهی اوقات بر روی داده-های زیست محیطی و داده های ساختار شیمیایی متمرکز شده اند. همکاری کتابخانه ای برای داده های کریستالوگرافی شیمی [۱۸] وجود دارد. در حالیکه کتابخانه های دانشگاهی تقریبا به همان اندازه که موسسات آموزشی دانشگاهی در آن ساکن هستند پایدار هستند (پایداری کامل و طولانی مدت ممکن است نداشته باشند)و مدل های بودجه از بخش از مأموریت های پژوهشی و آموزشی موسسات را پشتیبانی می کنند، بعید است که بار اضافی از داده ها در سطح بودجه فعلی مدیریت شود.

---

[۲۰] National science foundation

[۲۱] National Oceanic and Atmospheric Administration

[۲۲] National Center for Ecological Analysis

ناشران از قرن بیستم به تولید مجله علمی تسلط یافتند و برخی هم اکنون شروع به ارتباط داده ها با نشریات می کنند. در ادبیات زیست شناسی این ارتباط با بانک ژن به خوبی برقرار شده است. در حال حاضر بیشتر نشریات اطلاعات در پشت نمودار ها و آمار است که با ارزش است اما بسیاری از مسائل را حل نمی کند. زیرا قابلیت جستجوی اطلاعات محدوداست، و مکانیزمی برای ذخیره سازی اطلاعات در سطوح بالا وجود ندارد.[۱]

۳- ۲رویکردهای امیدوار کننده

خدمات و سازمان های ذکر شده و مانند آنها نیز در حال کار بر روی راه حل موانع استفاده از داده های موثر هستند که در بالا ذکر شده اند. در این قسمت برخی از راه حل ها و سازمان هایی که بر روی آنها کار می کنند را لیست می کند در حالی که راه حل های ذکر شده در اینجا جامع نیستند و در بعضی موارد ممکن است اثبات نگردیده باشد، آنها نمونه ای از فضای راه حل برای مشکل هستند. متاسفانه هیچ راه حل برای بهینه سازی حفظ و استفاده از داده ها وجود ندارد. با توجه به موانع بسیار استفاده از داده های بهینه، بسیاری از موسسات باید در ایجاد یک ساختار پاداش حرفه ای برای دانشمندان برای مشارکت با یکدیگر همکاری کنند. به اشتراک گذاری و حفاظت درازمدت داده ها باید منجر به موفقیت حرفه ای شود. دانشمندان در حال حاضر برای ارجاع به مقالات منتشر شده خود اعتبار می گیرند. اعتبار مشابه برای استفاده از داده ها نیاز به تغییر در جامعه شناسی علم است که در آن استناد به داده ها ارزش علمی داده می شود. صنایع چاپ و نشر با انتشار داده در کنار مقالات و در واقع ایجاد ارتباط بین داده و نشریه سعی در حل این موضوع داشته است. با این حال، محدودیت فضا، کنترل قالب و نمایه سازی اطلاعات، یک مشکل عمده باقی می ماند. مخازن سازمانی و انضباطی باید امکانات را فراهم کنند تا بتوانند مجموعه داده های مشابه بازگردانند. نهادهای استاندارد برای علوم می توانند روش هایی را برای فهرست ارجاعات داده ها در پایگاه های داده و نه فقط داده ها در نشریات [۱۸] ایجاد کنند.

در حال حاضر ناشرین و سرویس دهنده های نشانه معیار هایی دارند که با استناد به آنها داده های تولید شده توسط محققین و پروژه ها را ارزش گذاری میکنند و این ارزش گذاری به اعتبار محقق می افزاید اعتبار حرفه ای تولید کنندگان را تایید می کند.

باید منابع کافی برای انتخاب، حاشیه نویسی، حفظ و انتشار داده ها داشته باشد. کتابخانه ها یک راه حل واضح هستند، اما کتابخانه ها با محدودیت های مالی مواجه هستند[۱۸]اغلب مخازن انضباطی و مرجع از بودجه های پروژه ای استفاده میکنند این بودجه ها برای حفظ و نگهداری داده برای مدت ۳ تا ۵ سال داده میشود و بعد از آن یا بار دیگر باید تخصیص بودجه تمدید شود یا داده ها در دسته ی داده های تاریک قرار گیرند .

شرکت های بزرگ مانند googleو Microsoft و همچنین ناشران سنتی در سطح وسیعی(چند ۱۰۰ ترابایت)شروع کردند به ارائه اطلاعات تاریک پنهان در غبار یا در سایه ای که تا به امروز ذخیره کرده بودند. برای مجموعه دیتاهای کوچیک تر گوگل را منتشر کرد.

موسسات تجاری نیز در ساخت ابزارهای داده های علوم مولکولی شرکت کردند. مدل های مالی برای این کار هنوز روشن نیست. به رغم تمام این تلاش ها، مجموعه داده های تخصصی با کمبود قابلیت همکاری همچنان ادامه دارد. دانشگاه ها و مراکز داده ها شروع به ارائه برنامه های برای آموزش مراقبت و نگهداری از داده ها می کنند. آژانس فدرال ایالات متحده، موسسه خدمات موزه و کتابخانه در نهایت بودجه ای در نظر گرفت برای برگزاری کارگاه های پیشرفته ی آموزشی در این حوزه برای دانش جویان کارشناسی ارشد دانشگاها.[۱]

دانشگاهها ی پیشرو در این حوزه: دانشگاه ایلینوی ، دانشگاه کارولینای شمالی، و دانشگاه آریزونا و ...بودند که مرکز آموزش دیجیتال در بریتانیا آموزش های حرفه ای را فراهم کرده است و همکاری بین المللی در آموزش و پرورش از طریق گروه کاری[۲۳] (IDEA) آغاز شده است.

به برخی از موانع مالکیت فکری می توان با آموزش دانشمندان و کارکنان پشتیبانی آنها پرداخت. همان گونه که نگهداری دیتا تخصص حرفه ای است و هزینه به آن تخصیص داده میشود، ما باید مکانیزم ها و قوانین حقوقی را نیز در این حوزه در نظر بگیریم ، مهم است که دانشمندان تصمیمات آگاهانه در مورد کنترل حقوق مالکیت فکری را تصحیح کنند تا اثر مثبت بر روی علم به حداکثر برسد. فن آوری ها برای نگهداری دیتا آسان شده اند. سازمانهای دولتی و غیردولتی ابزارهای توسعه و ارزیابی در اختیار دارند. مخازن سازمانی و انضباطی درچهارچوب [۲۴]های رایج شروع به کار می کنند تا هزینه های توسعه را بین تعداد زیادی از کاربران به اشتراک بگذارند. موانع هنوز هم به اندازه کافی بالا هستند، با این حال اکثر دانشمندان به درستی اطلاعات خود را برای مدت طولانی مدیریت نمی کنند.

فرمت های فرا داده اجازه می دهد توضیحات داده ها یکپارچه شوند. زبان فرا داده اکولوژیکی [۲۵](EML) [۱۹] یک مثال در این زمینه است. هستی شناسی[۲۶] به تعریف روابط میان عناصر فردی مجموعه داده ها کمک می کند تا آنها را به یکدیگر متصل سازند. نمونه هایی از جمله:هستان شناسی ژن ها[۲۷]، هستان شناسی گیاهان[۲۸] چارچوب هایی

---

[۲۳] International Data Curation Education

[۲۴] framwork

[۲۵] Ecological Metadata Language

[۲۶] Ontologies

[۲۷] Gen Ontologies

[۲۸] Plant Ontologies

که برای به اشتراک گذاری داده ها استفاده میشود توسط ابزارها و نرم افزار های رایگان یا با هزینه ی کم بهبود پیدا کرده است.

این ابتکارات اطلاعاتی را که در غیر این صورت در پایگاه داده های انفرادی نامناسبی قرار می گرفت، در اختیار شما قرار می دهد. وب معنایی[۲۹] وعده داده است که در حل مسائل دسترسی به داده ها سهم قابل توجهی داشته باشد، اما دادههای دم بسیار وسیع و متنوع هستند.[۶]

۳-۳روشی دیگر برای روشن کردن داده های تیره

برای نشان دادن رویکرد ، ما یک نمونه اولیه را که به عنوان یک ویکی معنایی[۳۰] ساخته شده است توصیف میکنیم لینک داده جدید را وارد میکنیم . میتوان هر محتوای جدید ایجاد شده توسط کاربران را به عنوان داده های مرتبط منتشر کرد[۶].

گرچه دانشمندان در بسیاری از رشته ها اطلاعات را از طریق کاتالوگ به اشتراک می گذارند، تا دیگران بتوانند از این داده ها برای تجزیه و تحلیل و انتشارات (از جمله در نجوم، فیزیک و غیره) استفاده کنند، این الگو در محیط زیست خوب کار نکرده است. محیط زیست یک رشته علمی است که بسیاری از دانشمندان ابزارهای جمع آوری داده های خود را دارند و اغلب اطلاعات خود را برای سال های زیادی به یک مکان خاص نگهداری میکنند مقادیر زیاد داده ها بر روی سیستم های محلی هزاران دانشمند نشسته، اغلب به نام "داده های تاریک" عنوان میشود[۱].این مجموعه داده ها اغلب برای یک مکان یا پدیده بسیار خاص هستند، اما آنها توسط اکثریت قریب به اتفاق دانشمندان به نام "دمیدن طولانی علوم[۱۷]" توسعه می یابند. بعضی گزارش می دهند که کمتر از ۱٪ از داده ها در محیط زیست پس از تجزیه و تحلیل و نتایج منتشر شده است. اگرچه دانشمندان می خواهند داده ها را به اشتراک بگذارند، اغلب این کار را به دو دلیل اساسی انجام نمی دهند[۱].

۱. .برای بعضی از فعالان محیط زیست داده یک دارایی اولیه است
۲. اطلاعات در محیط زیست پیچیده هستند، در سطح بالایی از توزیع شدگی قرار دارند و به طور معمول برای پاسخ به سوالات محلی به دست آمده، و ارسال این داده ها به شیوه ای قابل کشف با دسترسی ساده به کار زیادی نیاز دارد

---

[۲۹] Semantic web

[۳۰] Semantic wiki

بسیاری از پروژه های فعلی در علوم زمین وابسته به دسترسی گسترده به داده ها در دم طولانی علم است. بسیاری از شبکه ها و طرح های رصدخانه ای مانند پروژه های زمین شناسی[۳۱] که با مشکلات اکوسیستم، منطقه ای، قاره ای و جهانی در ارتباطند. به عنوان مثال، برای درک چرخه کربن در آب، ادغام داده ها و تجزیه و تحلیل توسط دانشمندان، مطالعه رودخانه، دریاچه، اقیانوس و اکوسیستم های ساحلی نیاز است. تحقیقات انتقادی در بوم شناسی و علوم زمین تنها با تلفیق داده ها و مدل هایی از هزاران دانشمند د بسیاری از رشته ها (اقیانوس، زمین و علوم جوی) می تواند مورد توجه قرار گیرد. این پروژه ها نیاز به داده هایی دارند که می بایست به اشتراک گذاشته شوند، اما علاوه بر این، داده ها باید برای حمایت از اشتراک داده ها و همکاری های متقابل به طور همزمان پشتیبانی شوند. داده ها باید به طور آشکار در دسترس و با فرا داده ها به خوبی توضیح داده شود تا بتوان آن را جمع و یکپارچه کرد. [۶]

این کار با سه تکنیک مرتبط است:

۱.استانداردهای وب معنایی

۲.وب سایتهای مرتبط با اصول داده ها

۳.الگوهای محبوب وب برای رابط هایی مانند ویکی های معنی دار برای حاشیه نویسی و جمع آوری داده ها. کار با یک مرور کلی از رویکرد شروع می شود.به اشتراک گذاری داده های ارگانیک بر مبنای سه تکنیک مرتبط می شود:

۱. استانداردهای وب معنایی برای تعریف متادیتا های معنایی به صورت گسترده ای بر استانداردهای وب، از جمله استفاده از RDF برای تعریف انواع داده ها و خواص، که به کاربران اجازه می دهد تا دوباره از خواصی که قبلا در سیستم تعریف شده استفاده کنند و یا به راحتی خواص جدیدی را به سیستم و داده ها اضافه کنند.

۲. مجموعه داده ها و فرا داده ها مطابق با اصولی به هم مرتبط می شوند، نگهداری شده و منتشر میشوند. مخازن داده سنتی، داده ها را در پایگاه داده های مرکزی یا توزیعی بارگذاری می کنند، داده های مرتبط در این مخازن(هم اصل داده ها و هم فرا داده های مرتبط با آنها) به عنوان شی شبکه ای در اختیار برنامه های کاربردی وب ها قرار میگیرند مقادیر گسترده ای از اطلاعات مرتبط که به سرعت در حال رشد هستند در این فرمت منتشر میشوند. در حال حاضر این سیستم شامل مقادیر زیادی از مجموعه داده های مربوط به محیط زیست، مانند داده های جغرافیایی [۳۲]است

---

[۳۱] Earth-Cub

[۳۲] OpenStreetMap

۳. ویکی های معنایی به عنوان الگوهای محبوب وب برای رابط هاو دسترسی به تسهیل ایجاد ابزار ساده برای کاربردهای گسترده مانند تجسم، حاشیه نویسی و یکپارچه سازی داده ها ایجاد شده اند.ویکی های معنایی، ویکی های سنتی را تقویت می کنند تا پیوندهای بین عنوان ها و موضوعات هر صفحه با یک رابطه معنا دار شناخته شوند. همکاران نیز می توانند ساختار جدیدی از محتوا را با اضافه کردن خواص جدید ایجاد کنند.

رویکرد ما طراحی محیطی است که از دانشمندان پشتیبانی می کند تا فعالیت های زیر را انجام دهند:

•هر دانشمند می تواند وظایف مشارکتی را با تعریف سوالاتی که به مشارکت جامعه وسیع تر نیاز دارد تعریف کند

• هر دانشمند می تواند به این وظایف کمک کند، در صورت لزوم آنها را به یکسری زیر مسئله تقسیم کند و انواع مختلف داده ها را درخواست کند

• دانشمندان می توانند داده های مجموعه ای را به اشتراک بگذارند که به سادگی با اضافه کردن یک اشاره گر به مجموعه داده هایشان که به سیستم های محلی و تحت کنترل آنها متصل است امکان پذیر است.

• هر دانشمند می تواند متادیتا را به هر مجموعه داده اضافه کند، ویژگی های جدید فراداده را تعریف کند و یا خواص دیگری را که دیگران تعریف کرده اند استفاده کنند.

• هر دانشمند می تواند متادیتای مشخص شده برای هر مجموعه داده را تغییر دهد تا خواص مشابهی را که سایر مجموعه داده های مشابه استفاده می کنند، استفاده کند.

• هر دانشمند می تواند از هر مجموعه داده استفاده کند و باید نتایج تجزیه و تحلیل خود را با لینک های مناسب به مجموعه داده های اصلی که از آنها استفاده می کنند را ارسال کند.

سیستم از داده های اصلی پشتیبانی خواهد کرد:

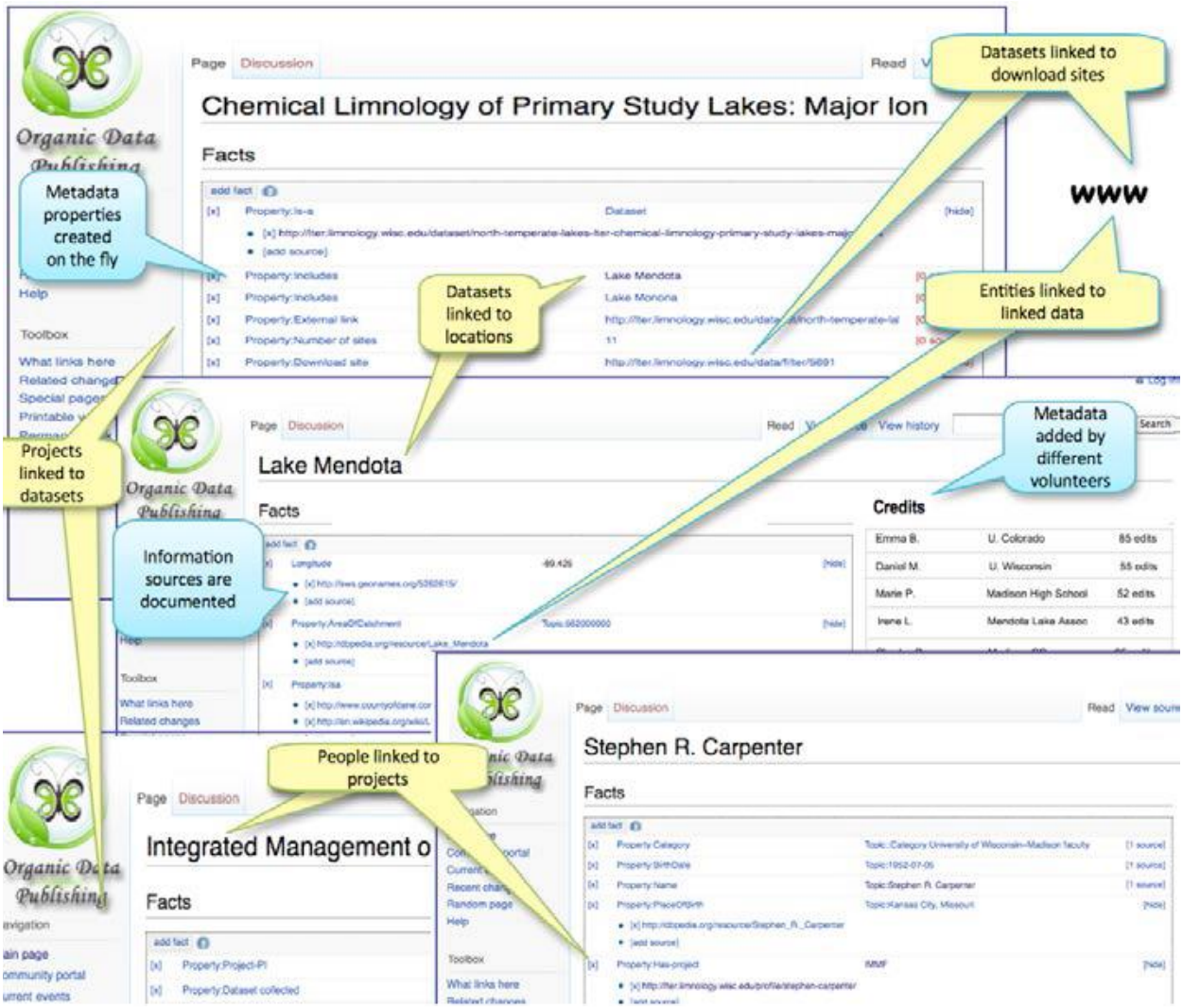
• اختصاص دادن اعتبار به هر دانشمند مستقل که تولید دیتا می کند.

• به کاربران اجازه داده میشود که محتوای جدید تولید کنند.

• انتشار هر گونه محتوای ایجاد شده توسط کاربران به عنوان داده های مرتبط شناخته می شود.

این بخش انتشار داده های ارگانی را از طریق یک نمونه اولیه که چهارچوب معنایی معنوی را گسترش می دهد، نشان می دهد. ویکی های رسانه ای معنایی بر روی نرم افزار MediaWiki محبوب ایجاد شده و آنها را گسترش می دهد تا کاربران بتوانند روابط معناشناختی را بیان کنند.

شکل ۳-۲ انواع مختلف موجودات را نشان می دهد که می توانند با خواص ساختاری مرتبط باشند. [۶]



شکل ۳- ۲ پنجره یک صفحه ویکی برای یک مجموعه داده ها[۶]

این شکل پنجره یک صفحه ویکی برای یک مجموعه داده ها همراه با متادیتا های توصیفی را نشان می دهد.

۳ پنجره در شکل است اطلاعات نهفته در پنجره ها بدین صورت است:

۱.مجموعه دیتا با متادیتا

۲.همان مجموعه دیتا با یک متا دیتای متفاوت

۳.دیگر دیتاهای ایجاد شده از آن دست

هر کسی می تواند ویکی را ویرایش کند، تمام ویژگی های متادیتا را اضافه کند، واژگان فراداده را گسترش دهد، و همه اطلاعات جمع آوری شده از طریق سایت به عنوان لینک داده منتشر می شود. همه ی نویسندگان هر صفحه در ابتدا تایید می شوند و پیوندی واضح با دانشمندانی که هر یک از مجموعه داده های اصلی را پشتیبانی می کنند، وجود دارد. این سیستم به همکاران اجازه می دهد تا به راحتی خواص معنایی ساختاری را برای توصیف محتویات ویکی ، تولید وبا استفاده از RDF استاندارد سازی کنند. هر صفحه ویکی یک شیء مورد علاقه را توضیح می دهد (به عنوان مثال یک مجموعه داده، یک پروژه) و دارای بخش " ویژگی های ساختار یافته" است که در آن نویسندگان می توانند خواص و مقادیر موضوع صفحه را مشخص کنند. هر مشارکت کننده می تواند خواص جدید را تعریف کند. هر سازنده می تواند یک ویژگی[۳۳] موجود را تغییر دهد یا آن را با یکی که در جاهای دیگر مورد استفاده قرار می گیرد جا به جا کند، استفاده از ویژگی هادر سراسر صفحات و در نتیجه در سراسر اشیاء نرمال است.

	Area of Catchment	Latitude	Longitude
Lake Casitas	100,000,000	34.392	-119.335
Lake Mendota	562,000,000	43.107	-89.425
Lake Monona			
Lake Wingra		43.053	-89.422

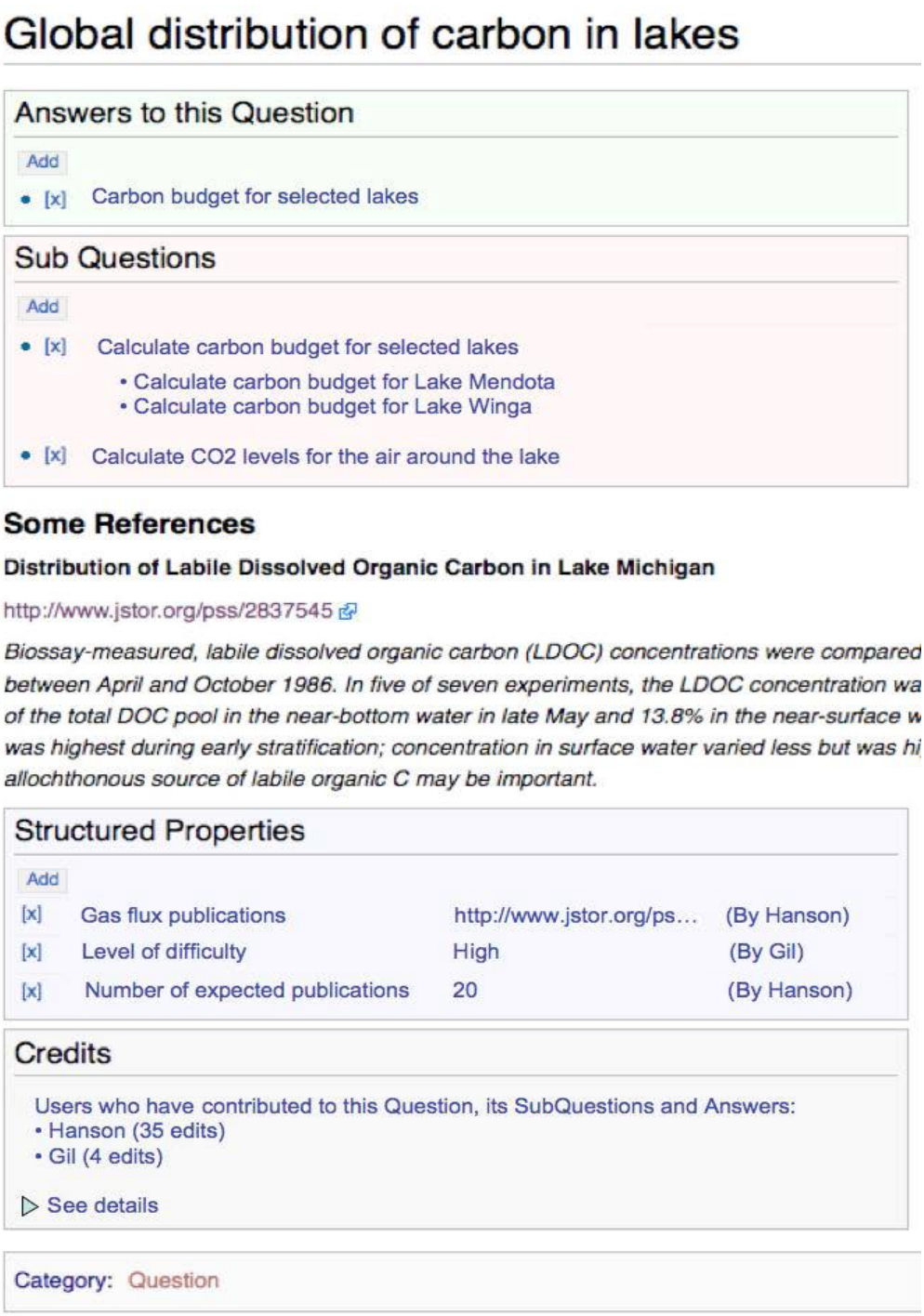
شکل ۳- ۳ ایجاد محتویات صفحات ویکی را از طریق نمایه ها[۶]

شکل ۳-۳ مثالی از چگونگی ایجاد محتویات صفحات ویکی را از طریق نمایه ها به صورت دینامیک نشان می دهد، در این حالت پرس و جو کاربران در حال دیدن سایت، بلافاصله در معرض اطلاعات گم شده قرار می گیرند و می توانند به آن کمک کنند و در یافتن داده مشارکت کنند.

[۳۳] properties

چارچوب دارای دسته بندی های پیش از تعریف صفحات است .ما تا کنون پنج دسته خاص را تعریف کرده ایم: سوال، پاسخ، داده، جریان کاری[۳۴] وجریان کاری در حال اجرا.

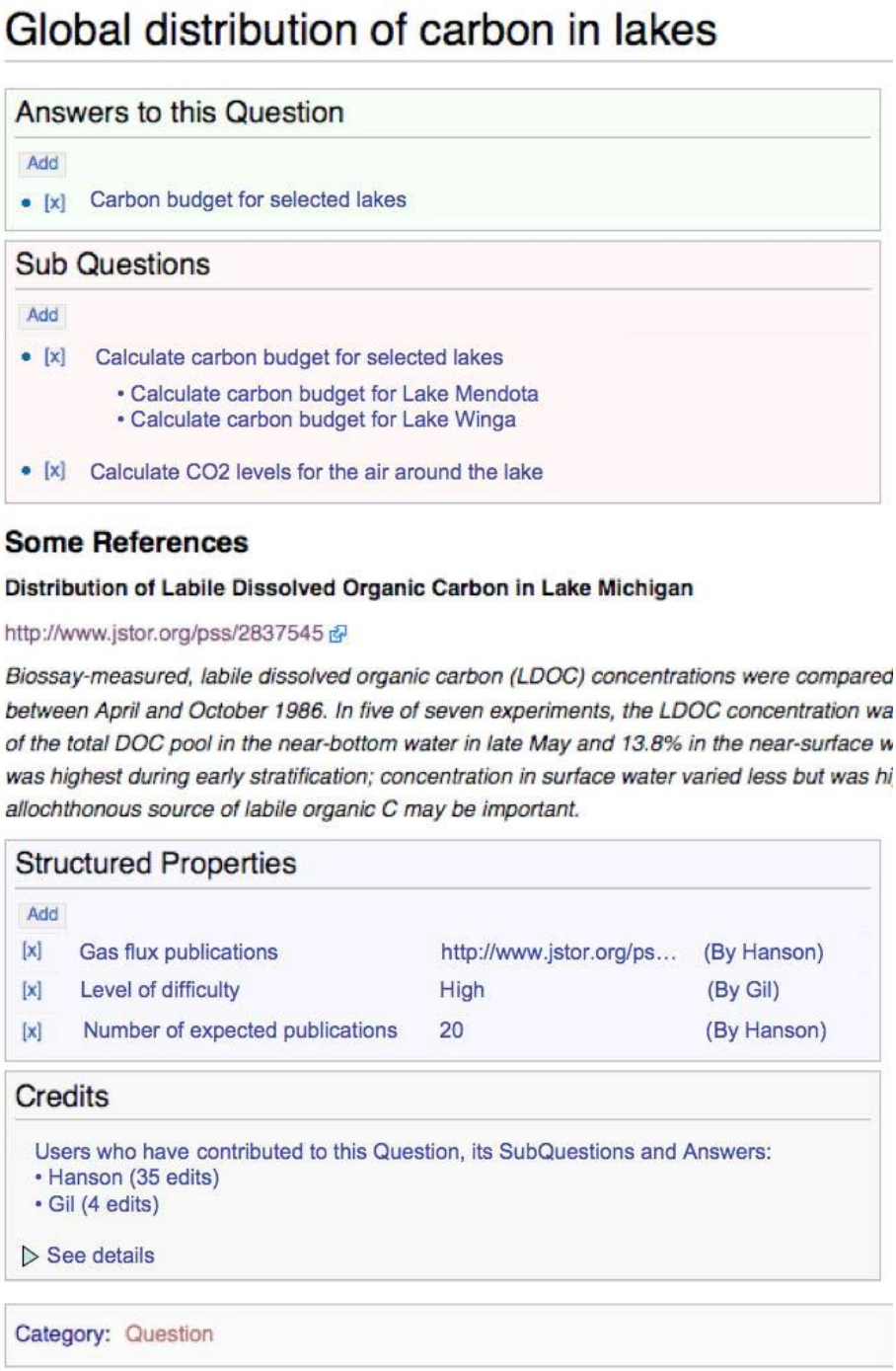
[۳۴] workflow



شکل ۳- ۴دسته بندی سوالات یک صفحه خاص را نشان میدهد[۶]

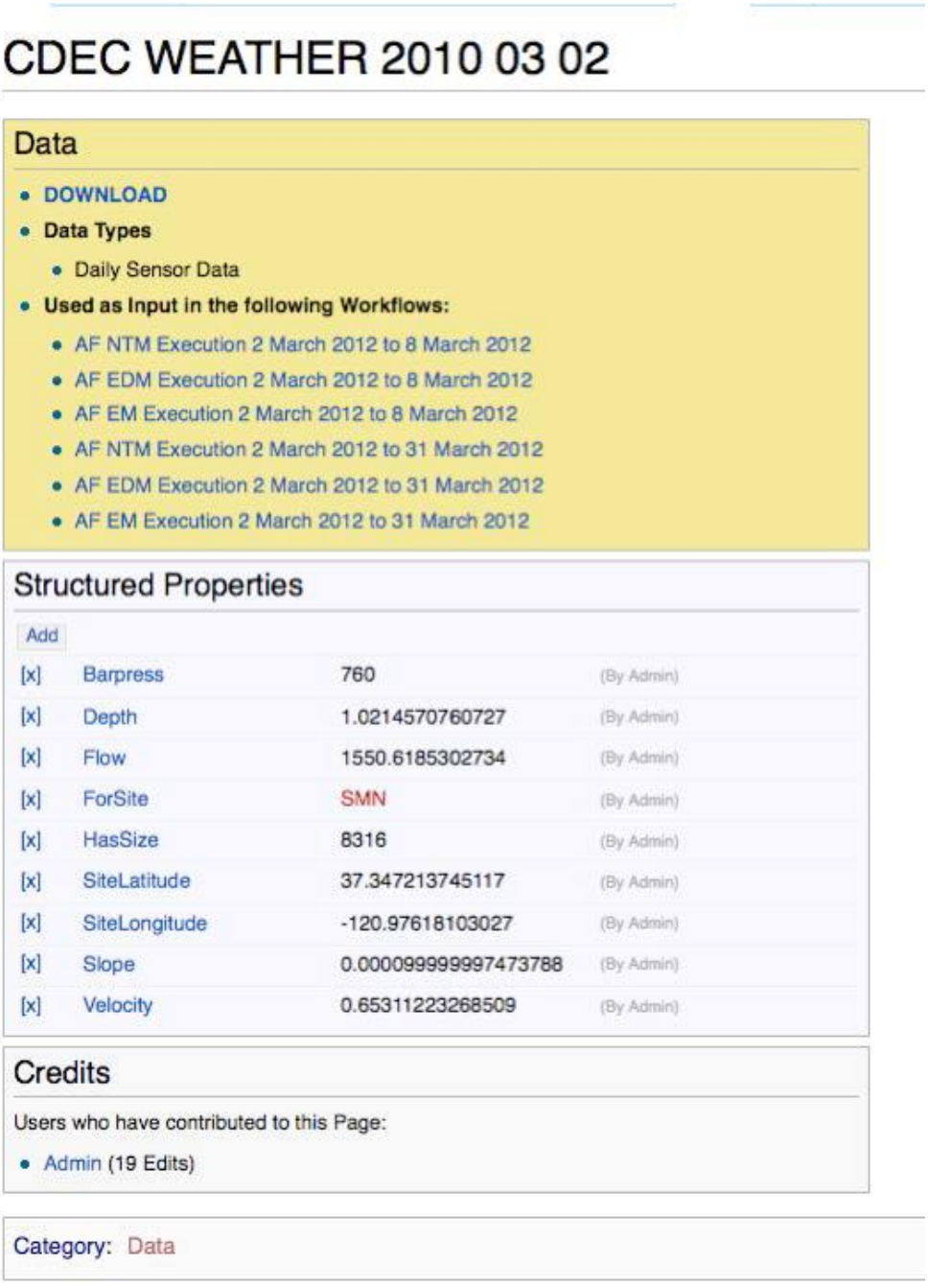
اینها صفحاتی هستند که منعکس کننده یک وظیفه یا زیر وظیفه هستند. آنها یک سری زیر سوال را دارند که به صفحاتی که پرسش ها در آنها دسته بندی شده نیز اشاره دارند.این زیر سوالات ممکن است منجر به درخواست یک مجموعه داده شود، همانطور که در نمونه ای که در شکل نشان داده شده است.

برخی از گردش کارها ممکن است طراحی شده و بعدا (پس از جمع آوری مجموعه داده های دلخواه.)اجرا شوند، هنگامی که سوال پاسخ داده می شود، کاربران می توانند پیج دیگری بادسته بندی از جواب ها راایجاد کنند که تمام یافته ها را خلاصه کند شاید صفحه و دسته بندی ایجاد شده شامل اشاره گرهایی به نشریات هم باشد.



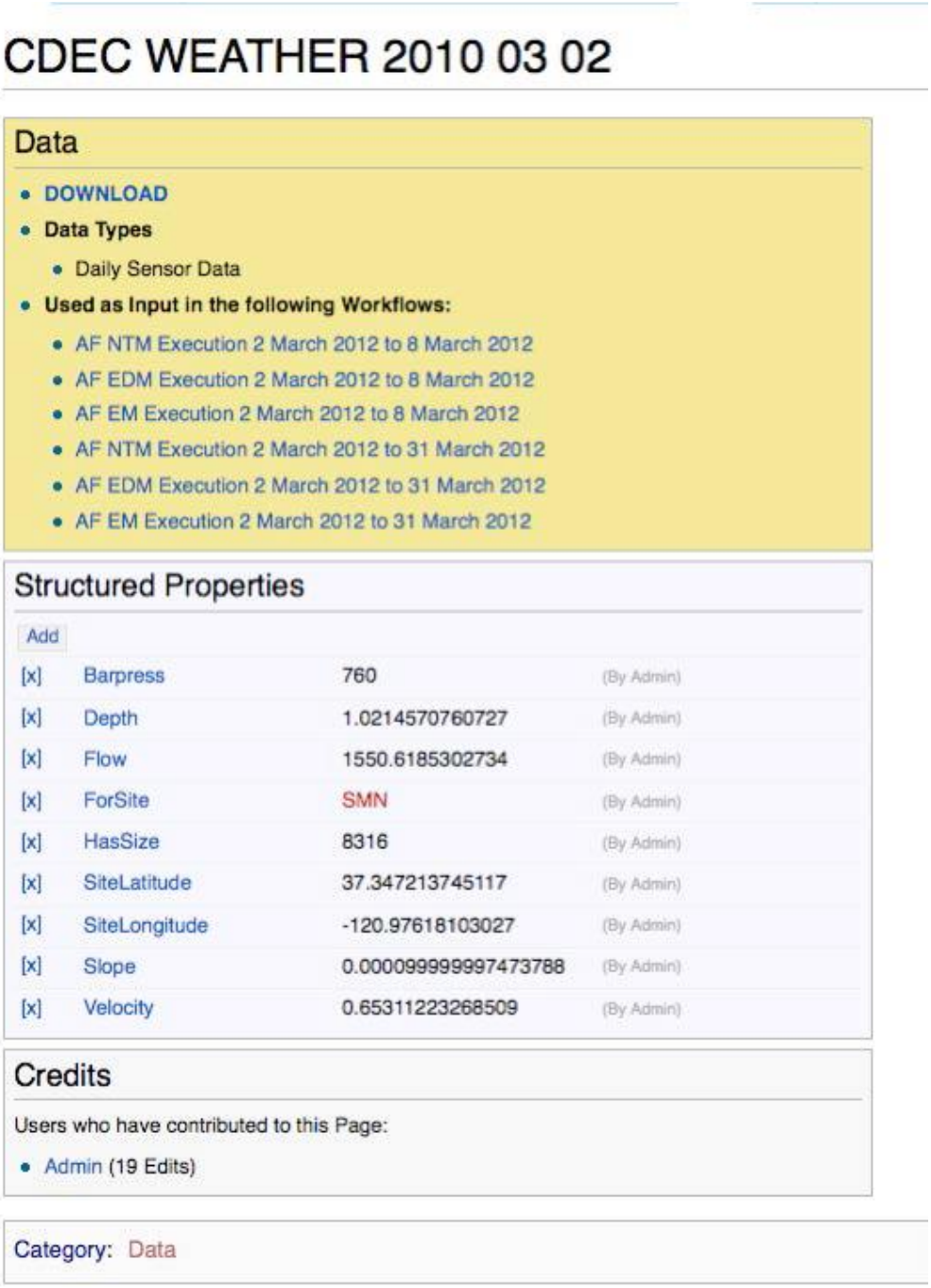
شکل ۳- ۵ دسته بندی وظایف و آدرس دهی زیر وظیفه ها[۶]

مانندهر صفحه دیگر، صفحات سوال می توانند ویژگی های ساختاری داشته باشند، و هر کدام به نویسنده آن اعتبار می دهند.



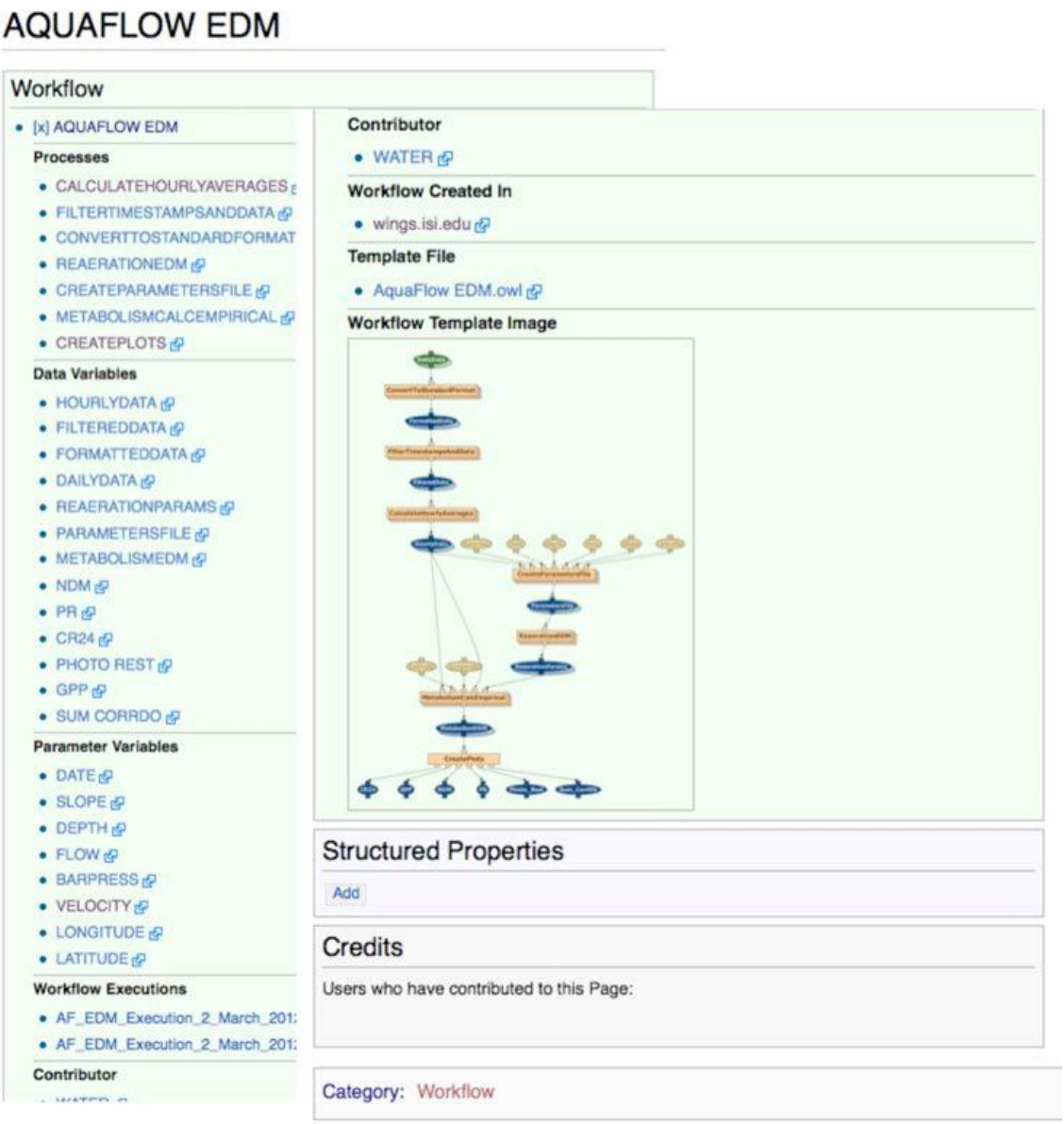
شکل ۳- ۶صفحات ویژه مخصوص دیتا ها[۶]

این صفحات نشان دهنده ی مجموعه داده ها هستند و میتوانند ویژگی های ساختار یافته داشته باشند .برخی از بخش های صفحه به صورت پویا از طریق پرس و جوها ایجاد میشوند.



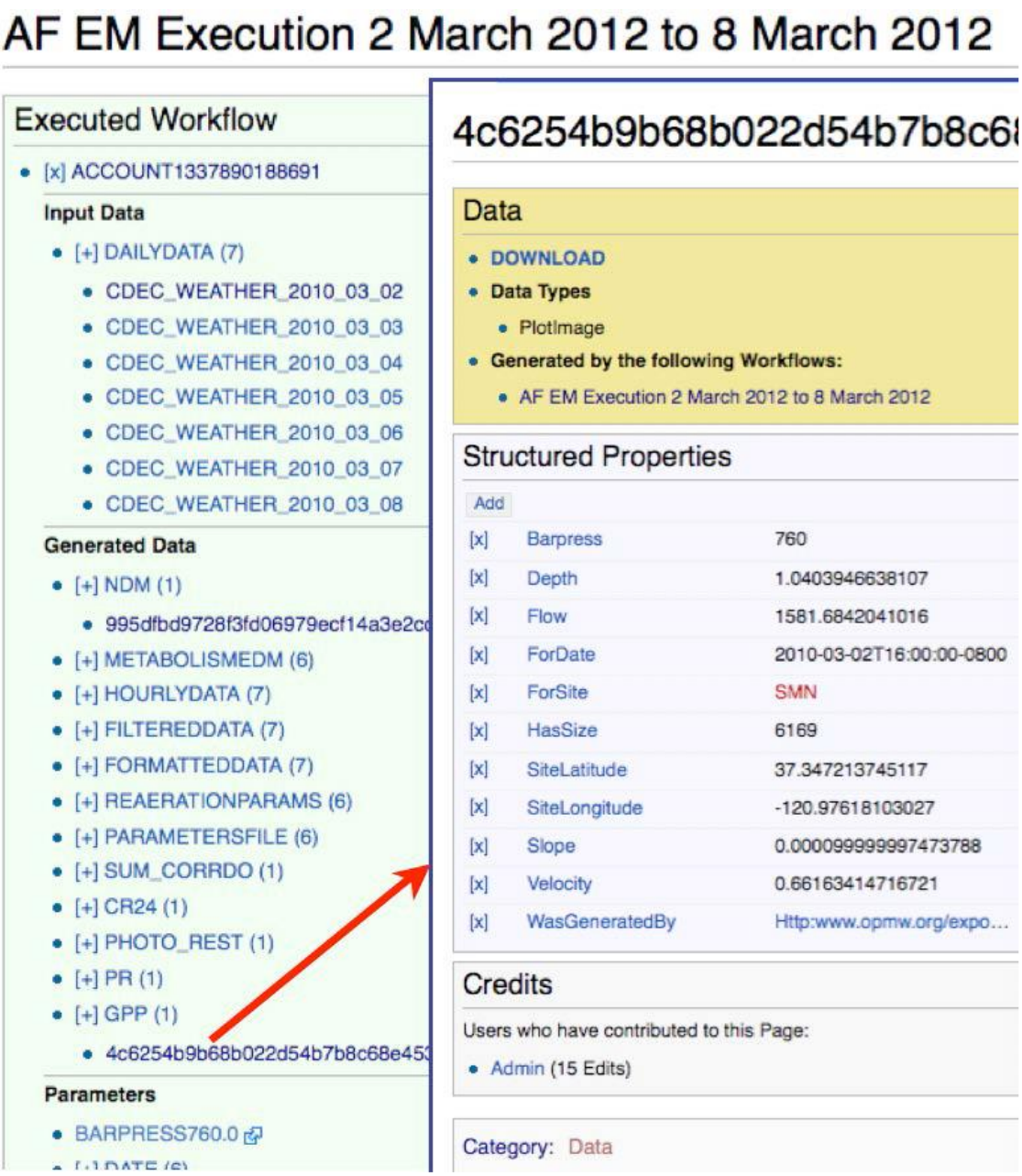
شکل ۳- ۷صفحات ویژه مخصوص دیتا ها[۶]

ویکی های معنایی میتوانند بطور پویا برای پیج ها محتوا تولید کنند.به عنوان مثال جریان های داده ای که از داده-های خاص به عنوان ورودی استفاده میکنند.



شکل ۳- ۸ک جریان کاری را نشان می دهد[۶]

مشارکت برای پاسخگویی به سوالات علمی جهانی، یک انگیزه بزرگ برای مشارکت دانشمندان است. پاسخ دادن به این سؤالات، اهداف کلانی است که به کارکنان نیاز دارد تا کارهای مختلفی را انجام دهند مانند تقسیم سوالات سطح بالا به وظایف کوچکتر، به اشتراک گذاشتن مجموعه داده ها، توصیف ویژگی های داده، تهیه آنها، اجرای مدل ها و غیره.



شکل ۳- ۹ قسمت جریان کاری در حال اجرا را نشان می دهد[۶]

شکل بالا صفحه ای که برای یکی از محصولات داده های گردش کار تولید شده است. اجرای یک گردش کار را می توان به صفحه سوال مناسب مرتبط کرد .تکنولوژی جریان کاری ویک سری پرتکل های استاندارد تعبیه شده- اند تا دانشمندان را قادر سازند به توصیف فرایند های تحلیلی، که نحوه ی دستیابی به داده های جدید از داده های

خام را ثبت میکنند. گردش کارها و نتایج آنها نیز می تواند به صورت دستی توسط کاربران اضافه شود، مثلا اگر مراحل با دست یا از طریق اسکریپت انجام شود. برای مخاطب غیر فنی باید ابزار هایی ساده در نظر گرفته شود.

همچنان این نمونه اولیه را گسترش داده شده است تا رویکرد به اشتراک گذاری داده های خام را نشان دهد. کار با جامعه EarthCube برای شناسایی نیازهای اضافی از دانشمندان بخش بعدی کار است. باید به کسانی که در این حوزه تحقیق میکنند و داده تولید میکنند اعتبار داده شود و به نحوی از آنها تقدیر شود.

داده های کمی را می توان با استفاده از ابزار سیستم جمع آوری کرد. ما می توانیم از معیارهای جمع آوری داده های ویکی که در انجام مطالعات رفتار کاربر و رشد محتوا (مثلا تعداد ویرایش ها در هر کاربر) استفاده میشود استفاده کنیم. می توان معیارهای خاصی برای ویکی های معنایی (مثلا تعداد ویژگی های ساخت یافته تعریف شده) داشت.علاوه بر این ارزیابی های معمول سنتی ویکی، می توان آن را با معیار های دیگر هم سنجید. مانند تعداد مجموعه داده های جمع آوری شده و تعداد مجموعه داده های ایجاد شده در حین فرایند نرمال سازی. یکی دیگر از جنبه های جدید درگیر در ارزیابی سیستم، در مورد تجزیه وظیفه، مشارکت کار و انجام وظیفه ای است که مورد توجه همکاری قبلی در مشارکت قبلی قرار نگرفته است.

چهار بعد مهم ارزیابی را که مورد توجه قرار گرفته اند، مشخص می کنیم: مشارکت[۳۵]، همکاری[۳۶]، همگرایی[۳۷] و دستیابی به جامعه[۳۸].

معیارهای مشارکت می توانند مورد استفاده قرار گیرند که نشان دهنده دخالت کاربران از جامعه است. ما می توانیم تخمینی از اندازه جامعه را به عنوان تعداد کل کاربران منحصر به فرد که همیشه از سایت بازدید می کنند، ایجاد کنیم. سپس سیستم می تواند تعداد کل کاربران را که صفحات را ویرایش می کنند و محتوایی را به سایت اضافه کنند، تعداد کل مجموعه داده ها و تعداد کل ویرایش ها را به صورت جمعی وبرای هر کاربر، جمع آوری کند.بعلاوه معیار های مشارکتی میتوانند با توجه به ویژگی های ساختار یافته ی ویکی های معنایی، جمع آوری داده را عهده دار شوند.[۲۰]

---

[۳۵] participation

[۳۶] collaboration

[۳۷] convergence

[۳۸] Archivement of the community

.معیار همکاری بیان میکند چطور فعالیت ها ی کاربرانی که روی یک موضوع خاص و یک صفحه از ویکی با یکدیگر همکاری میکند با یکدیگر هم پوشانی دارد. دیتا ها میتواند جمع آوری شود برا اساس تعدا کاربرانی که آن پیج خاص در رابطه با ان موضوع خاص را ویرایش کردند. در واقع دیتا ی مرتبط با هر عنوان خاص از پیج ها و مراجعی براداشته میشود که در فواصل کمتری به روز رسانی و ویرایش شده اند و لینک بیشتری به انها وجود دارد (یعنی بیشتر مورد علاقه و توجه افراد متخصص یا محققین در آن حوزه ی خاص بوده اند.)و درصد توزیع شدگی بالاتری دارند که این نشان می دهد افراد بیشتری در حل آن مسئله خاص و زیر مسئله هایش با یکدیگر همکاری داشته اند.معیارهای همگرایی نشان می دهد که جامعه چگونه خصوصیات ساختاری را به عنوان ابرداده به مجموعه داده های متنوع اضافه می کند. این معیار شامل تعداد ویژگی های معمول در دسترسی به دیتا ست هایی است که برای آن وظیفه یاجریان کاری مورد نیاز است تعداد افراد منحصر به فردی که که هر ویژگی را به روز رسانی میکنندتعداد ویژگی های معنایی منسوخی که با ویژگی های دیگر جایگزین میشوند. این باعث تکامل خواص معنایی در طول زمان می شود. معیار دستیابی پیشرفت به این گونه است که سیستم می تواند معیارهای مربوط به مقدار وظایف و زیر وظایف را ایجاد کند مثلا میزان جمع آوری داده ها و صفحات گردش کار ایجاد شده در ارتباط با وظایف، میزان فعالیت کاربر در ارتباط باهر وظیفه و ....هرکدام معیارهای خاصی دارند. [۲۲]

یکی دیگر از راه حل های موجود تولید یک مقاله داده است. یک مقاله داده یک نشریه مجله است که هدف اصلی آن توصیف داده ها است تا گزارش تحقیقاتی به همین ترتیب، آن حاوی اطلاعات مربوط به داده هاست، نه فرضیه ها و استدلالات و حمایت از این فرضیه ها بر اساس داده ها، همانطور که در یک مقاله پژوهشی متعارف یافت می شود.

اهداف آن :

۱. تهیه یک نشر مجله قابل نقل و گواهی که اعتبار علمی را برای ناشران داده

۲. توصیف داده ها در یک فرم ساختار یافته انسانی قابل خواندن؛ وانتقال انها به جامعه علمی با توجه به نیاز جامعه

مقاله های داده سندهایی هستند که دیتا ست های منتشر شده ی مقالات را توصیف میکنند این مقالات همیشه به دیتا ست های منتشر شده وابسته هستند و بدون آنها معنی ندارند.و این پیوند ها url,doi هستندکه باید درون مقاله ی دیتا قرار بگیرند.

یکسری از داده ها هستند که در بایگانی ها نگهداری میشوند برای تولید متادیتا برای این داده ها باید یکسری اطلاعات اولیه راجع به آنها داشت این اطلاعات شامل پیونداین داده ها به مقالات معتبری است که یکبار در مجلات و ژورنال ها به چاپ رسیده است.

راه حل های گفته شده در راستای ایجاد یک نظام ساختار یافته برای تولید دیتا های ساختار یافته و روشن است که با سرمایه گذاری روی آنها می توان از بروز مشکلات جلوگیری کرد. برای استفاده از داده های تاریک حال حاضر نیز ابزار هایی معرفی شده که با تکیه بر الگوریتم های تکاملی و روش های یادگیری ماشین و شبکه عصبی ابزار هایی تولید شده تا بتواند در بعضی حوزه های خاص داده ها را ساختار یافته و روشن کند تا استفاده از آنها میسر شود یکی از این ابزار ها DeepDiveمی باشد. این ابزار در سال ۲۰۱۶ معرفی شد[۲۱] که با یک معماری نرم افزاری و سخت افزاری خاص و با استفاده از روش های استخراج ویژگی سعی در روشن کردن داده های تاریک داشت.و دیگری[۲۳] HITACHIکه باز بعنوان ابزاری بیزینسی طراحی و معرفی شد که امکان اتصال به برنامه- های کاربردی و صفحات وب را داشت و داده های تولید شده را دسته بندی و سازماندهی میکرد.روش های معرفی شده کماکان دارای مشکلات زیادی است و بسیار جای کار دارد.

۳-۴ خلاصه

این فصل رویکرد های مقابله با داده های تاریک را از دو منظر مورد نقد و بررسی قرار داد که یکی اجتناب از تولید داده های تیره و دیگری تلاش برای روشن کردن داده های تیره است. در ادامه راه حل هایی در این دو حوزه ارائه شد و سعی شد تا موانع مختلف در سر راه این چالش جدید مورد بررسی قرار بگیرد هر چند که برای رسیدن به هدف نهایی که یافتن راهی مطمئن برای روشن کردن داده های تیره است راهی طولانی در پیش داریم.

فصل ۴: خلاصه و نتیجه‌گیری

داده ها اساس روش های علمی هستن در حالی که اکثریت داده ها در شرکت های بزرگ علمی به خوبی سرپرستی می شود، زیرساخت های علمی کمی برای حمایت از ذخیره سازی و استفاده مجدد از داده های ایجاد شده توسط پروژه های کوچک وجود دارد. برای به حداکثر رساندن بازده سرمایه گذاری در تحقیقات علمی ما باید زیرساخت های علمی را از طریق موسسات موجود مانند کتابخانه ها و موزه ها که به طور سنتی نگهبانان بهره وری علمی هستند، توسعه دهیم. ما نیاز به توسعه فن آوری هایی داریم که برای دانشمندان هزینه می کند تا مدارک و داده های خود را در این مخازن ذخیره کنند. ما همچنین به ابزارهایی نیاز داریم که اطلاعات را از این مخازن جستجو و بازیابی کنند. ما نیاز به آموزش نسل جدیدی از مدیران خروجی علمی ما داریم که در فن آوری مناسب رایانه آموزش دیده اند و از علم و جامعه شناسی علم قدردانی می کنند. ما بیشتر نیازمند طرح های آموزشی جدید و انگیزه هایی هستیم که نسل بعدی دانشمندان دانش خود را برای تصمیم گیری های آگاهانه در مورد استفاده گسترده تر از داده ها و تاثیر گسترده تر تحقیقاتشان به دانش آموزان می دهند.

بسیاری از دانشمندان اطلاعات خود را به دلیل هزینه و عدم انگیزه روش های سنتی برای به اشتراک نمی گذارند. ما روش جدیدی برای به اشتراک گذاری داده ها ارائه دادیم با ویژگی های زیر:

۱) مجموعه داده ها را بطور مستقیم به سوالات علمی مرتبط میکند.
۲) با توانمند کردن هر دانشمند برای مشارکت ایجاد فرا داده ها بار اشتراک داده را کاهش می دهد.
۳) ردیابی و اعطای اعتبار به همه افرادی که در ایجاد دیتا و فرا دیتا ها نقش داشتند.

پس از گذر از چالش کلان داده ها و داده های تاریک راجع به مشکلات و معضلات موجود در این حوزه سخن گفتیم همانطور که مشخص است حوزه کلان داده به خصوص بخش داده های تاریک حوزه ای تازه و قابل بحث است و تا به امروز خیلی به آن پرداخته نشده است. در ادامه اشاره کردیم که در این حوزه چند چالش وجود دارد جمع آوری داده ها ، نگهداری و سازماندهی که هر کدام چالش های مربوط به خود را داشت.

از دو زاویه می توان به این حوزه نگاه پژوهشی داشت یکی ساخت و طراحی سیستمی که از ابتدا داده ها را ساخت_ یافته جمع آوری، نگهداری وسازماندهی کند دوم ساخت و طراحی سیستمی که بتواند داده های تاریک موجود را روشن کند.در گزارش به هر دوی این دو نگاه پرداخته شده است در ادامه در نظر داریم از منظر نگاه دوم داده های تاریک را رصد کنیم و برای به روشنایی آوردن داده های تاریک چاره ای بیاندیشیم.

مراجع

[١] Heidorn, P.B. “Shedding Light on the Dark Data in the Long Tail of Science.” Library Trends, Vol. ٥٧, No. ٢, Fall ٢٠٠٨.

[

[٢] Coles, S. J., Frey, J. G., Hursthouse, M. B., Light, M. E., Milsted, A. J., Carr, L., De Roure, D.,Gutteridge, C., Mills, H. R., Meacham, K., Surridge, M., Lyon, E., Heery, R., Duke, M., & Day, M. (٢٠٠٦). An E-Science environment for service crystallography-from submission to dissemination. *Journal of Chemical Information and Modeling* ٤٦(٣): ١٠٠٦–١٠١٦

[٣] Darwin Core. (n.d.). Retrieved October ٣١, ٢٠٠٨, from http://wiki.tdwg.org/twiki/bin/view/DarwinCore/

[٤] DataNet. (n.d.). Retrieved October ٣١, ٢٠٠٨, from http://www.nsf.gov/funding/pgm_summ jsp?pims_id.٥٠٣١٤١

[٥] https://www.kdnuggets.com/٥/٠٣/٢٠١٨-things-big-data.html

[٦] Organic Data Publishing: A Novel Approach to ScientificData SharingYolanda Gil and Varun RatnakarInformation Sciences Institute and Department of Computer ScienceUniversity of Southern California gil@isi.eduPaul C. Hanson
Center for Limnology

[٧] J. Betteridge, A. Carlson, S. A. Hong, E. R. H. Jr., E. L. M.Law, T. M. Mitchell, and S. H. Wang. Toward never endinglanguage learning. In Learning by Reading and Learning toRead, Papers from the ٢٠٠٩ AAAI Spring Symposium, Technical Report SS-٠٩-٠٧, Stanford, California, USA, March ٢٣-٢٥, ٢٠٠٩, pages ١-٢, ٢٠٠٩.

[٨] S. Brin. Extracting patterns and relations from the worldwide web. In The World Wide Web and Databases, International Workshop WebDB'٩٨, Valencia, Spain, March ٢٧-٢٨, ١٩٩٨, Selected Papers, pages ١٧٢{١٨٣, ١٩٩٨.

[٩] R. C. Bunescu and R. J. Mooney. Learning to extractrelations from the web using minimal supervision. In ACL٢٠٠٧, Proceedings of the ٤٥th Annual Meeting of theAssociation for Computational Linguistics,

. June ٢٣-٣٠, ٢٠٠٧, Prague, Czech Republic, ٢٠٠٧

[١٠] Y. Chen and D. Z. Wang. Knowledge expansion overprobabilistic knowledge bases. In International Conferenceon Management of Data, SIGMOD ٢٠١٤, Snowbird, UT, USA, June ٢٢-٢٧, ٢٠١٤, pages ٦٤٩-٦٦٠, ٢٠١٤.

[١١] X. L. Dong, E. Gabrilovich, G. Heitz, W. Horn, K. Murphy,S. Sun, and W. Zhang. From data fusion to knowledge fusion. PVLDB, ٧(١٠):٨٨١-٨٩٢, ٢٠١٤.

[١٢] http://www.ibmbigdatahub.com/infographic/four-vs-big-data

[١٣] S. Krause, H. Li, H. Uszkoreit, and F. Xu. Large-scalelearning of relation-extraction rules with distant supervision from the web. In ISWC, ٢٠١٢

[١٤] https://timoelliott.com/blog/٠٣/٢٠١٥/what-is-big-data-discovery.html

[١٥] http://odintext.com/blog/shedding-light-on-dark-data-how-to-get-started/

[١٦] Reichman, O.J., Jones, M.B., and M.P. Schildhauer. “Challenges and Opportunities of Open Data in Ecology.” Science, Vol. ٣٣١ no. ٦٠١٨ pp. ٧٠٣-٧٠٥, February ٢٠١١, DOI: ١٠,١١٢٦/science.٣٣١,٦٠١٨,٦٩٢.

[١٧] Rocca RA, Magoon G, Reynolds DF, Krahn T, Tilroe VO, et al. “Discovery of Western European R١b١a٢ Y Chromosome Variants in ١٠٠٠ Genomes Project Data: An Online Community Approach.” PLoS ONE ٧(٧), ٢٠١٢.

[١٨] Science, ٢٠١١, Special Issue on Challenges and Opportunities. Vol. ٣٣١ no. ٦٠١٨ pp. ٦٩٢-٦٩٣, February ٢٠١١, DOI: ١٠,١١٢٦/science.٣٣١,٦٠١٨,٦٩٢.

[١٩] Effect of the combination of white and red LED lighting during incubationon layer, broiler, and Pekin duck hatchabilityG. S. Archer,*,١ D. Jeffrey,† and Z. Tucker†*Department of Poultry Science, Texas A&M University; College Station, TX, ٧٧٨٤٥, USA; and †Maple Leaf Farms, Inc., Leesburg, Indiana ٤٦٥٣٨, USA

[٢٠] J. Liu, S. J. Wright, C. R_e, V. Bittorf, and S. Sridhar. Anasynchronous parallel stochastic coordinate descentalgorithm. In Proceedings of the ٣١th InternationalConference on Machine Learning, ICML ٢٠١٤, Beijing, China, ٢١-٢٦ June ٢٠١٤, pages ٤٦٩-٤٧٧, ٢٠١٤.

[۲۱] E. K. Mallory et al. Large-scale extraction of gene interactions from full text literature using deepdive.
Bioinformatics, ۲۰۱۵.

[۲۲] M. R. Anderson, D. Antenucci, V. Bittorf, M. Burgess, M. J.Cafarella, A. Kumar, F. Niu, Y. Park, C. R_e, and C. Zhang.Brainwash: A data system for feature engineering. In CIDR۲۰۱۳, Sixth Biennial Conference on Innovative DataSystems Research, Asilomar, CA, USA, January ٦-٩, ۲۰۱۳Online Proceedings, ۲۰۱۳.

[۲۳] Garijo, D., and Gil, Y. "A New Approach for Publishing Workflows: Abstractions, Standards, and Linked Data". In Proc. WORKS'۱۱, Seatle, WA, ۲۰۱۱

[۲۴] Kamar, E., Hacker, S. and E. Horvitz. "Combining Human and Machine Intelligence in Large-scale Crowdsourcing," AAMAS ۲۰۱۲, Valencia, Spain, June ۲۰۱۲.